Model-Based RL for Mean-Field Games is not Statistically Harder than Single-Agent RL

Jiawei Huang [1] **Niao He** [1] **Andreas Krause** [1]

Abstract

We study the sample complexity of reinforcement learning (RL) in Mean-Field Games (MFGs) with model-based function approximation that requires strategic exploration to find a Nash Equilibrium policy. We introduce the Partial Model-Based Eluder Dimension (P-MBED), a more effective notion to characterize the model class complexity. Notably, P-MBED measures the complexity of the single-agent model class converted from the given mean-field model class, and potentially, can be exponentially lower than the MBED proposed by Huang et al. (2023). We contribute a model elimination algorithm featuring a novel exploration strategy and establish sample complexity results polynomial w.r.t. P-MBED. Crucially, our results reveal that, under the basic realizability and Lipschitz continuity assumptions, *learning Nash Equilibrium in MFGs is no more statistically challenging than solving a logarithmic number of single-agent RL problems*. We further extend our results to Multi-Type MFGs, generalizing from conventional MFGs and involving multiple types of agents. This extension implies statistical tractability of a broader class of Markov Games through the efficacy of mean-field approximation. Finally, inspired by our theoretical algorithm, we present a heuristic approach with improved computational efficiency and empirically demonstrate its effectiveness.

1. Introduction

Multi-Agent Reinforcement Learning (MARL) has excelled in modeling cooperative and competitive interactions among agents in unknown environments. However, the well-known "curse of multi-agency" poses a challenge in equilibrium solving for MARL systems with large populations. Yet, for MARL systems with symmetric agents, such as human crowds or road traffic, one can leverage such special structure by employing mean-field approximation, leading to the RL for Mean-Field Games (MFGs) setting (Huang et al., 2006; Lasry & Lions, 2007). Notably, MFGs offer a promising framework where the complexity of learning Nash Equilibrium (NE) needs not depend on the number of agents (Laurière et al., 2022). It has found successful applications in various domains, including financial markets (Cardaliaguet & Lehalle, 2018), economics (Gomes & Pimentel) and energy management (Djehiche et al., 2016).

Similar to single-agent RL (Jiang et al., 2017; Jin et al., 2018), for MFGs, one of the most important questions is to understand how many samples are required to explore the unknown environment and solve the equilibrium, a.k.a. the *sample complexity*. Given the complex dynamics of mean-field systems and high cost of generating samples from large population, designing strategic exploration methods for sample-efficient learning becomes imperative.

Existing works on learning MFGs primarily focus on model-free approaches such as Q-learning (Anahtarci et al., 2023; Guo et al., 2019), policy gradient (Subramanian & Mahajan, 2019; Yardim et al., 2022), fictitious play (Perrin et al., 2020; Xie et al., 2021), etc. Several recent works further extend these model-free approaches with value function approximation to handle large state-action space (Mao et al., 2022; Zhang et al., 2023). However, their sample complexity results ubiquitously rely on strong structural assumptions such as contractivity (Guo et al., 2019) or monotonicity (Perolat et al., 2021). Their methods, moreover, are usually specialized and lack generalizability, leaving an open challenge of efficiently exploring mean-field systems *without those structural assumptions*.

To address this gap, Huang et al. (2023) establish general sample complexity results for model-based RL in MFGs[1]. They introduce a complexity measure known as Model-Based Eluder Dimension (MBED) to characterize the com-

[1]Department of Computer Science, ETH Zurich. Correspondence to: Jiawei Huang <jiawei.huang@inf.ethz.ch>.

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

[1]Model-based RL has also been explored in Mean-Field Control (MFC) setting, where all the agents are cooperative (Huang et al., 2023; Pasztor et al., 2021).

plexity of the model function class. Their algorithm, under basic realizability and Lipschitz continuity assumptions, achieves a sample complexity upper bound polynomial w.r.t. MBED. However, as we will show in Prop. 3.4, even for the tabular setting, MBED can be exponential in the number of states in the worst case. This observation, coupled with the tractability of tabular MFGs under additional structural assumptions (Guo et al., 2019; Perolat et al., 2021), prompts a fundamental question:

Is learning MFGs statistically harder than single-agent RL in general?

In this paper, we provide a definitive answer to this question. Our main contributions are summarized as follows:

- In Sec. 3, we introduce a novel complexity measure for any given mean-field model class $\mathcal{M}$, called *Partial Model-Based Eluder Dimension (P-MBED)*. P-MBED represents the complexity of the single-agent model class derived from $\mathcal{M}$ after (adversarially) fixing the state density for the transition functions in $\mathcal{M}$. We show that P-MBED can be significantly lower than MBED (Huang et al., 2023). For example, in the tabular setting, P-MBED is always bounded by the number of states and actions, yielding an exponential improvement over MBED.

- In Sec. 4, we propose a model elimination algorithm capable of exploring the mean-field system and returning an approximate NE policy with sample complexity polynomial w.r.t. P-MBED. From the algorithmic perspective, our results indicate that under the basic realizability and Lipschitz assumptions, *learning MFGs is no more statistically challenging than solving* $\log|\mathcal{M}|$ *single-agent RL problems*. As a direct implication, the sample complexity of tabular MFGs only polynomially depends on the number of states, actions, horizon and $\log|\mathcal{M}|$. This is the first result indicating that learning tabular MFGs is provably sample-efficient in general, even without the contractivity or monotonicity assumptions.

- In Sec 6, we design a heuristic algorithm with improved computational efficiency building upon our insights in theory. We evaluate it in a synthetic linear MFGs setting and validate its effectiveness.

As a substantial extension, we further examine the sample complexity of more general MFGs with heterogeneous population, specifically Multi-Type MFGs (MT-MFGs) (Ghosh & Aggarwal, 2020; Perolat et al., 2021; Subramanian et al., 2020). MT-MFGs comprise multiple types of agents with distinct transition models, reward functions or even state-action spaces. MT-MFGs have stronger capacity in modeling the diversity of agents, while being more tractable than general Markov Games[2]. However, the fundamental sample complexity in the setting remains largely unexplored. Our additional contribution includes:

- In Sec. 5, we show that finding the NE in an MT-MFG is equivalent to finding the NE in a lifted MFG with constraints on policies. Building on this insight, we establish the first sample complexity upper bound for learning MT-MFGs. Our results identify statistical tractability of a broad class of MARL systems, potentially offering new insights to the sample complexity analysis for solving NE in general Markov Games.

1.1. Related Work

Within the abundant literature on single-agent RL and MFGs, below we focus primarily on sample complexity results for solving these problems in unknown environments. We defer additional related works to Appx. B.

Single-Agent RL When the number of states and actions is extremely large, sample complexity bounds derived for tabular RL (Auer et al., 2008; Azar et al., 2017; Jin et al., 2018) become vacuous. Instead, function approximation is usually considered, where a model or value function class containing the true model or optimal value functions is available, and the sample complexity is governed by the complexity of the function classes (Agarwal et al., 2020; Du et al., 2021; Foster et al., 2021; Jiang et al., 2017; Jin et al., 2020; 2021a; Sun et al., 2019). Compared with single-agent RL, the main challenge in MFGs is the additional dependence on density in transition and reward functions, especially that the density space is continuous. Although our P-MBED is inspired by the eluder dimension in the single-agent setting (Levy et al., 2022; Osband & Van Roy, 2014; Russo & Van Roy, 2013), it is a novel complexity notion in characterizing the sample efficiency of RL in MFGs.

Mean-Field Games Most existing results for learning MFGs primarily focus on tabular setting and model-free approaches (Cui & Koeppl, 2021; Elie et al., 2020; Guo et al., 2019; Xie et al., 2021), where strong structural assumptions, such as contractivity (Guo et al., 2019), monotonicity and density independent transition (Perrin et al., 2020), or non-vanishing regularization (Yardim et al., 2022) are usually required. In contrast, we focus on addressing the fundamental exploration challenge for general MFGs. Mishra et al. (2020) study non-stationary MFG without strong structural

[2]The general Markov Games (MGs) framework considers individually distinct agents. However, this generality comes with challenges. Existing results in MGs are restricted in learning (Coarse) Correlated Equilibria (Bai et al., 2020; Daskalakis et al., 2023; Jin et al., 2021b) and the sample complexity in function approximation setting may still depend on the number of agents (Cui et al., 2023; Wang et al., 2023). MT-MFGs can be regarded an intermediary between standard MFGs and general MGs.

assumptions, but their algorithm is inefficient and no sample complexity results were provided. Beyond the tabular setting, Huang et al. (2023) is the most related to us. However, as implied by our results in this paper, their sample complexity bound are suboptimal.

Multi-Type Mean-Field Games Subramanian et al. (2020) study more general multi-type cases, but they consider the transition model depending on action density instead of state density. Besides, the multi-type setting has been investigated in special cases, such as LQR (Moon & Başar, 2018; uz Zaman et al., 2023), and leader-follower structures (Vasal & Berry, 2022). Ghosh & Aggarwal (2020) is the most related to us. However, they consider the discounted stationary setting and assume the state density is fixed, while ours is more challenging since we need to keep tracking the evolution of state density. Perolat et al. (2021) also consider the multi-type setting, but they require the monotonicity assumption. Moreover, they only provide asymptotic rates without sample complexity guarantees.

2. Background

Notations Throughout the paper, we will use standard big-oh notations $O(\cdot), \Omega(\cdot), \Theta(\cdot)$, and notations such as $\widetilde{O}(\cdot)$ to (partially) suppress logarithmic factors. In Appx. A, we list all the frequently used notations in this paper.

2.1. Mean-Field Games

Mean-Field Markov Decision Process We consider the finite-horizon non-stationary Mean-Field MDP (MF-MDP) $M := (\mu_1, \mathcal{S}, \mathcal{A}, H, \mathbb{P}_M, r)$, where μ_1 is the known initial state distribution; $\mathcal{S} = (\mathcal{S}_1 = ... = \mathcal{S}_H)$ and $\mathcal{A} = (\mathcal{A}_1 = ... = \mathcal{A}_H)$ are the state and action spaces, which are discrete but can be arbitrarily large; $\mathbb{P}_M := \{\mathbb{P}_{M,h}\}_{h\in[H]}$ with $\mathbb{P}_{M,h} : \mathcal{S}_h \times \mathcal{A}_h \times \Delta(\mathcal{S}_h) \to \Delta(\mathcal{S}_{h+1})$ is the transition function and $r := \{r_h\}_{h\in[H]}$ with $r_h : \mathcal{S}_h \times \mathcal{A}_h \times \Delta(\mathcal{S}_h) \to [0, \frac{1}{H}]$ is the deterministic reward function, where $\Delta(\mathcal{X})$ denotes the probability measure over $\mathcal{X}$. We use $\Pi := \{\pi := \{\pi_h\}_{h\in[H]} | \pi_h : \mathcal{S}_h \to \Delta(\mathcal{A}_h)\}$ to denote the policy class including all non-stationary Markovian policies, and we only focus on policies in Π. Given $\pi \in \Pi$ and initial density $\mu^\pi_{M,1} := \mu_1$, the state density $\mu^\pi_M := \{\mu^\pi_{M,h}\}_{h\in[H]}$ evolves according to $\mu^\pi_{M,h+1} = \Gamma^\pi_{M,h}(\mu^\pi_{M,h}), h \in [H]$ where $\Gamma^\pi_{M,h}(\mu_h)(\cdot) := \sum_{s_h,a_h} \mu_h(s_h)\pi(a_h|s_h)\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu_h)$.

Given any $\pi, \widetilde{\pi} \in \Pi$, we use $\mathbb{E}_{\widetilde{\pi},M(\pi)}[\cdot]$ to denote the expectation over trajectories generated by executing policy $\widetilde{\pi}$ while fixing the transitions and rewards to $\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^\pi_{M,h})$, $r_h(\cdot,\cdot,\mu^\pi_{M,h})$. These trajectories can be interpreted as the observations of a deviated agent taking $\widetilde{\pi}$ while all the others take π. Besides, we define $V^{\widetilde{\pi}}_{M,h}(\cdot;\mu^\pi_M) := \mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}, \mu^\pi_{M,h'})|s_h = \cdot]$ to be the value function at step h if the agent deploys policy $\widetilde{\pi}$ in model M conditioning on π, and define $J_M(\widetilde{\pi};\pi) := \mathbb{E}_{s_1\sim\mu_1}[V^{\widetilde{\pi}}_{M,1}(s_1;\mu^\pi_M)]$ to be the total return of policy $\widetilde{\pi}$ conditioning on π. The Nash Equilibrium (NE) π^{NE}_M of model M is defined to be the policy s.t. no agent tends to deviate, i.e., $\forall\widetilde{\pi} \in \Pi$, $J_M(\widetilde{\pi};\pi^{\text{NE}}_M) \le J_M(\pi^{\text{NE}}_M;\pi^{\text{NE}}_M)$. We denote $\mathcal{E}^{\text{NE}}_M(\pi) := \max_{\widetilde{\pi}} \Delta_M(\widetilde{\pi},\pi)$ to be the NE-Gap, where $\Delta_M(\widetilde{\pi},\pi) := J_M(\widetilde{\pi};\pi) - J_M(\pi;\pi)$.

Model-Based Setting In our model-based setting, the learner can get access to a transition function class $\mathcal{M} \subset \{\{\mathbb{P}_{M,h}\}_{h\in[H]} | \forall h, \mathbb{P}_{M,h} : \mathcal{S}_h \times \mathcal{A}_h \times \Delta(\mathcal{S}_h) \to \Delta(\mathcal{S}_{h+1})\}$ to approximate the true model M^*. We assume the reward function r is known. In Appx. C, we provide informal discussion about how to extend our results to the setting when r is unknown. Our main objective is to find an ε-approximate NE $\widehat{\pi}^{\text{NE}}_{M^*}$, satisfying $\mathcal{E}^{\text{NE}}_{M^*}(\widehat{\pi}^{\text{NE}}_{M^*}) \le \varepsilon$. Same as Huang et al. (2023), we only make two basic assumptions about the function class $\mathcal{M}$: realizability and Lipschitz continuity.

Assumption A (Realizability)**.** $M^* \in \mathcal{M}$.

Assumption B (Lipschitz Continuity)**.** For any $M \in \mathcal{M}$, and arbitrary policies $\pi, \widetilde{\pi} \in \Pi$, $\forall h, s_h, a_h$, we have:

$$\begin{aligned}
&\|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^\pi_{M,h}) - \mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^{\widetilde{\pi}}_{M,h})\|_1 \\
&\qquad \le L_T \|\mu^\pi_{M,h} - \mu^{\widetilde{\pi}}_{M,h}\|_1, \\
&|r_h(s_h, a_h, \mu^\pi_{M,h}) - r_h(s_h, a_h, \mu^{\widetilde{\pi}}_{M,h})\|_1 \\
&\qquad \le L_r \|\mu^\pi_{M,h} - \mu^{\widetilde{\pi}}_{M,h}\|_1.
\end{aligned}$$

Note that our Assump. B only requires Lipschitz continuity on feasible densities. In contrast, contractivity assumes L_r and L_T are sufficiently small (Guo et al., 2019; Yardim et al., 2022), and prior works considering monotonicity (Perolat et al., 2021; Zhang et al., 2023) usually assume the transition is independent w.r.t. density, i.e., $L_T = 0$.

We consider the same trajectory sampling model as Huang et al. (2023), which is much weaker than the generative model assumptions requiring trajectories conditioning on arbitrary state densities in most MFGs literatures (Anahtarci et al., 2023; Guo et al., 2019; Perrin et al., 2020).

Definition 2.1. The sampling model can be queried with arbitrary $\widetilde{\pi}, \pi \in \Pi$, and return a trajectory by executing $\widetilde{\pi}$ while transition and reward functions are fixed to $\mathbb{P}_{M^*,h}(\cdot|\cdot,\cdot,\mu^\pi_{M^*,h})$ and $r_h(\cdot,\cdot,\mu^\pi_{M^*,h})$ for all h.

MFGs and N-Player Symmetric Anonymous Games MFGs can be regarded as the limit of Symmetric Anonymous Games (SAGs) when the number of agents N approaches infinity (Guo et al., 2019; Yardim et al., 2022). As explained in (Huang et al., 2023), the sampling model (Def. 2.1) is reasonable for N-player SAGs with central controllers, which can manipulate all the agents' policies.

Given a SAG, it is known that the NE of its MFG approximation is a $O(N^{-1/2})$-approximate NE for the SAG (Yardim et al., 2024), if all the agents execute that same NE policy. In this way, one may interpret our setting as centralized training with decentralized execution.

2.2. Multi-Type Mean-Field Games

Multi-Type Mean-Field MDP A finite horizon non-stationary Multi-Type (or Multi-Group) MF-MDP $\boldsymbol{M}$ with W types of agents can be described by a collection of tuples $\boldsymbol{M} := \{(\mu_1^w, \mathcal{S}^w, \mathcal{A}^w, H, \mathbb{P}_{\boldsymbol{M}}^w, r^w)_{w\in[W]}\}$, where we use w in superscription to distinguish the initial state distribution, state-action spaces, the transition and reward functions in different groups. Besides, for any w, the transition and reward functions depend on densities in all types. More concretely, we have $\mathbb{P}_{\boldsymbol{M}}^w := \{\mathbb{P}_{\boldsymbol{M},h}^w\}_{h\in[H]}$ with $\mathbb{P}_{\boldsymbol{M},h}^w : \mathcal{S}_h^w \times \mathcal{A}_h^w \times \Delta(\mathcal{S}_h^1) \times ... \times \Delta(\mathcal{S}_h^W) \to \Delta(\mathcal{S}_{h+1}^w)$ and $r^w := \{r_h^w\}_{h\in[H]}$ with $r_h^w : \mathcal{S}_h^w \times \mathcal{A}_h^w \times \Delta(\mathcal{S}_h^1) \times ... \times \Delta(\mathcal{S}_h^W) \to [0, \frac{1}{H}]$. For each type of agents, we consider the Markovian policies $\Pi^w := \{\pi^w := \{\pi_h^w\}_{h\in[H]} | \forall h,\ \pi_h^w : \mathcal{S}_h^w \to \Delta(\mathcal{A}_h^w)\}$, and use $\boldsymbol{\Pi} := \{\boldsymbol{\pi} := \{\pi^w\}_{w\in[W]} | \forall w \in [W], \pi^w \in \Pi^w\}$ to denote the collection of policies for all types. For the function approximation setting, we assume W function classes $\mathcal{M}^1, ..., \mathcal{M}^W$ are available, where $\forall w \in [W]$, $\mathcal{M}^w \subset \{\{\mathbb{P}_h^w\}_{h\in[H]} | \forall h \in [H],\ \mathbb{P}_h^w : \mathcal{S}_h^w \times \mathcal{A}_h^w \times \Delta(\mathcal{S}_h^1) \times ... \times \Delta(\mathcal{S}_h^W) \to \Delta(\mathcal{S}_h^w)\}$ is used to approximate the transition function for the w-th group. The MT-MFG function class $\boldsymbol{\mathcal{M}}$ is then defined by $\boldsymbol{\mathcal{M}} \leftarrow \{\boldsymbol{M} := M^1 \times ... \times M^W | \forall w \in [W],\ M^w \in \mathcal{M}^w\}$, which we use to approximate the true model $\boldsymbol{M}^*$.

For the lack of space, we defer the definitions of value functions, Nash Equilibrium, and other related details to Appx. G.1. For the assumptions in MT-MFG setting, we defer to Appx. G.3.

3. Partial Model-Based Eluder Dimension

In the function approximation setting, the exploration challenge is related to the complexity of the function classes. In this section, we introduce new notions to characterize the complexity of model function class for MFGs and its extension to Multi-Type MFGs setting. The proofs and additional discussions can be found in Appx. D.

Inspired by the Eluder dimension of single-agent value function classes (Jin et al., 2021a; Russo & Van Roy, 2013) and mean-field model function classes (Huang et al., 2023), similarly, we use the length of independent sequences to characterize the complexity of function classes. In Def. 3.1, we first introduce the definition of standard ε-independence in previous Eluder dimension literature, to highlight the difference from our partial ε-independence. Although we only consider the l_1-distance here, similar discussion can be generalized to other distances, e.g., the Hellinger distance.

Definition 3.1 (ε-Independence; (Huang et al., 2023))**.** Given $\mathcal{M}$ and a data sequence $\{(s_h^i, a_h^i, \mu_h^i)\}_{i=1}^n \subset \mathcal{S}_h \times \mathcal{A}_h \times \Delta(\mathcal{S}_h)$, we say (s_h, a_h, μ_h) is ε-independent of $\{(s_h^i, a_h^i, \mu_h^i)\}_{i=1}^n$ w.r.t. $\mathcal{M}$ if there exists $M, \widetilde{M} \in \mathcal{M}$ such that $\sum_{i=1}^n \|\mathbb{P}_{M,h}(\cdot|s_h^i, a_h^i, \mu_h^i) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h^i, a_h^i, \mu_h^i)\|_1^2 \le \varepsilon^2$ but $\|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu_h) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu_h)\|_1 > \varepsilon$. We call $\{(s_h^i, a_h^i, \mu_h^i)\}_{i=1}^n$ an ε-independent sequence w.r.t. $\mathcal{M}$ (at step h) if for any $i \in [n]$, (s_h^i, a_h^i, μ_h^i) is ε-independent w.r.t. $\{(s_h^t, a_h^t, \mu_h^t)\}_{t=1}^{i-1}$.

Definition 3.2 (Partial ε-Independence)**.** Given $\mathcal{M}$, a mapping $\nu_h : \mathcal{M} \to \Delta(\mathcal{S}_h)$, and a data sequence $\{(s_h^i, a_h^i)\}_{i=1}^n \subset \mathcal{S}_h \times \mathcal{A}_h$, we say (s_h, a_h) is partially ε-independent of $\{(s_h^i, a_h^i)\}_{i=1}^n \subset \mathcal{S}_h \times \mathcal{A}_h$ w.r.t. $\mathcal{M}$ and ν_h, if there exists $M, \widetilde{M} \in \mathcal{M}$, s.t. $\sum_{i=1}^n \|\mathbb{P}_{M,h}(\cdot|s_h^i, a_h^i, \nu_h(M)) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h^i, a_h^i, \nu_h(\widetilde{M}))\|_1^2 \le \varepsilon^2$ but $\|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \nu_h(M)) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \nu_h(\widetilde{M}))\|_1 > \varepsilon$. We call $\{(s_h^i, a_h^i)\}_{i=1}^n$ a partially ε-independent sequence w.r.t. $\mathcal{M}$ and ν_h (at step h) if for any $i \in [n]$, (s_h^i, a_h^i) is partially ε-independent on $\{(s_h^t, a_h^t)\}_{t=1}^{i-1}$.

Intuitively, a partially ε-independent sequence of $\mathcal{M}$ is an independent sequence w.r.t. the function class converted from $\mathcal{M}$ by using some mapping ν_h to "partially" fix the input (the density part) for each function in $\mathcal{M}$. We use $\dim_{\mathrm{E}|\nu_h}(\mathcal{M}, \varepsilon)$ to denote the length of the longest partially ε-independent sequence w.r.t. $\mathcal{M}$ and ν_h (at step h). Now, we are ready to define the Partial-MBED.

Definition 3.3 (Partial MBED)**.** Given a model class $\mathcal{M}$, and a policy π, we define the mapping ν_h^π: $\forall M \in \mathcal{M}$, $\nu_h^\pi(M) := \mu_{M,h}^\pi$. The P-MBED of $\mathcal{M}$ is defined by: $\dim_{\mathrm{PE}}(\mathcal{M}, \varepsilon) := \max_{h\in[H]} \max_\pi \dim_{\mathrm{E}|\nu_h^\pi}(\mathcal{M}, \varepsilon)$.

By definition, P-MBED can be interpreted as the complexity of the *single-agent model class* converted from the Mean-Field model class $\mathcal{M}$ by *partially (adversarially) fixing the density* of the functions' input. In fact, different choices of ν in Def. 3.2 may lead to different notions of complexity. In our main text, we stick to the choice in Def. 3.3, but in Appx. D.1, we discuss an alternative choice of ν, its induced P-MBED and associated properties.

Next, we take the tabular setting as an example, and show that P-MBED of any function class for tabular MFGs can be controlled by $|\mathcal{S}||\mathcal{A}|$, while MBED (Huang et al., 2023) can be exponential in $|\mathcal{S}|$ in the worst case. This is reasonable given the single-agent nature of P-MBED.

Proposition 3.4. *(Tabular Setting) For any $\mathcal{M}$ and $\varepsilon > 0$, $\dim_{\mathrm{PE}}(\mathcal{M}, \varepsilon) \le |\mathcal{S}||\mathcal{A}|$, while there exists a concrete example of $\mathcal{M}$ such that $\dim_{\mathrm{E}}(\mathcal{M}, \varepsilon) = \Omega(\exp(|\mathcal{S}|))$.*

Below we provide the linear mean-field model classes with decomposable transition functions as another exam-

ple. When the transition is independent w.r.t. density μ (i.e. $G(\mu)$ is constant), the linear MFGs reduce to the single-agent linear MDP (Jin et al., 2020). As we can see, the P-MBED of the model class of linear MFGs is only related to the dimension of the state-action feature, which matches the complexity of their single-agent correspondence.

Proposition 3.5 (Linear MFGs; Informal version of Prop. D.4)**.** *Consider the model class:* $\mathcal{M}_\Psi := \{\mathbb{P}_\psi | \mathbb{P}_\psi(\cdot|s,a,\mu) := \phi(s,a)^\top G(\mu)\psi(s'); \psi \in \Psi\}$*, with known feature* $\phi(\cdot,\cdot) \in \mathbb{R}^{\widetilde{d}}$*,* $G(\cdot) \in \mathbb{R}^{\widetilde{d}\times d}$*, and a next-state feature class* Ψ *satisfying some normalization conditions. Then* $\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon) = \widetilde{O}(\widetilde{d})$.

Similarly, for model classes in Multi-Type MFGs setting, we can define the Multi-Type P-MBED generalized from $\dim_{\mathrm{PE}}$ in MFGs, which we denote as $\dim_{\mathrm{MTPE}}$. We defer its formal definition to Appx. D.3. Likewise, $\dim_{\mathrm{MTPE}}$ can be regarded as the complexity measure for a collection of W single-agent model classes converted from $\boldsymbol{\mathcal{M}}$. In the tabular case (resp. Prop. D.11), we have $\dim_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}},\varepsilon') = \widetilde{O}(\sum_{w\in[W]}|\mathcal{S}^w||\mathcal{A}^w|)$, and in linear MT-MFG setting with decomposable transitions (resp. Prop. D.12), $\dim_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}},\varepsilon') = \widetilde{O}(\sum_{w\in[W]} d^w)$ where $\{d^w\}_{w\in[W]}$ are the dimensions of the state-action features.

4. Sample Efficiency of Learning in MFGs

In this section, we show that the sample complexity of learning NE in MFGs is indeed governed by our new complexity notion P-MBED. We highlight our main algorithm and sample complexity results in Sec. 4.1, and then explain details in the algorithm design and technical novelty in Sec. 4.2. The missing details and proofs for results in this section are deferred to Appx. F.

4.1. Main Algorithm and Highlight of Main Results

Before proceeding to the algorithms, we first introduce several useful notions. Given a reference policy π, we denote $d(M,\widetilde{M}|\pi) := \max_{\widetilde{\pi}} \; d^{\widetilde{\pi}}(M,\widetilde{M}|\pi) \vee d^{\widetilde{\pi}}(\widetilde{M},M,|\pi)$ as the conditional model distance between M and $\widetilde{M}$, where $d^{\widetilde{\pi}}(M,\widetilde{M}|\pi) := \mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^H \|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{M},h}(\cdot|\cdot,\cdot,\mu^\pi_{\widetilde{M},h})\|_1]$. Given a model class $\mathcal{M}'$, any $M \in \mathcal{M}'$, and any policy π, we define the ε_0-neighborhood of M in $\mathcal{M}'$ to be: $\mathcal{B}^{\varepsilon_0}_\pi(M;\mathcal{M}') := \{M' \in \mathcal{M}' | d(M,M'|\pi) \le \varepsilon_0\}$. The "Central Model" (abbr. CM) in $\mathcal{M}'$ w.r.t. policy π is defined to be the model with the most number of neighbors: $M^{\varepsilon_0}_{\mathrm{Ctr}}(\pi;\mathcal{M}') := \arg\max_{M\in\mathcal{M}'} |\mathcal{B}^{\varepsilon_0}_\pi(M;\mathcal{M}')|$. Besides, when ε_0 and $\mathcal{M}'$ is clear from the context, we will use M^π_{Ctr} as a short note of $M^{\varepsilon_0}_{\mathrm{Ctr}}(\pi;\mathcal{M}')$. Lastly, $\forall \pi,\pi' \in \Pi$, we define $d_{\infty,1}(\pi,\pi') := \max_{h,s_h} \|\pi(\cdot|s_h) - \pi'(\cdot|s_h)\|_1$.

We provide our main algorithm in Alg. 1. The basic idea is to find a sequence of "reference policies" (π^k or $\pi^{\mathrm{NE},k}_{\mathrm{Br}}$, $k = 1,2,...$) and run the model elimination steps (Alg. 2 as `ModelElim`) to gradually remove models in $\mathcal{M}$ that distinct from M^* conditioning on these reference policies, until find an approximate NE. Next, we highlight our main results and its implications.

Theorem 4.1. *[Informal version of Thm. F.7] Under Assump. A and B, with appropriate hyperparameter choices, w.p.* $1-\delta$*, Alg. 1 terminates at some* $k \le \log_2|\mathcal{M}|+1$ *and returns an* ε*-NE of* M^* *after consuming at most*

$$\widetilde{O}\left(\frac{H^7}{\varepsilon^2}(1+L_rH)^2 \dim_{\mathrm{PE}}(\mathcal{M},\varepsilon') \log^3 \frac{|\mathcal{M}|}{\delta}\right)$$

trajectories, where $\varepsilon' = O(\frac{\varepsilon}{H^3(1+L_rH)(1+L_T)^H})$.

Model-Based RL for MFGs is not Statistically Harder than Single-Agent RL As we will explain more in the next section, `ModelElim` only needs to be a *single-agent model elimination subroutine*, and it is the only step consuming samples. Therefore, Thm. 4.1 suggests that the sample complexity of learning MFGs can be characterized by a $O(\log|\mathcal{M}|)$ number of single-agent model elimination subproblems, whose learning complexity is controlled by P-MBED. As a result, the total sample complexity only scales with P-MBED and the log-covering number of $\mathcal{M}$.

Based on the discussion in Sec. 3, we can expect for many model classes with low P-MBED (e.g. tabular setting Prop. 3.4, linear setting Prop. 3.5), learning MFGs is provable sample-efficient. In particular, for tabular MFGs where $\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon') \le |\mathcal{S}||\mathcal{A}|$, our result yields a sample complexity with polynomial dependence on $|\mathcal{S}|,|\mathcal{A}|,H$, which implies that *tabular MFGs are provably efficient in general* if considering the model-based function approximation, even without assuming contractivity or monotonicity that are often required in existing works (Guo et al., 2019; Perrin et al., 2020; Yardim et al., 2022). Compared with recent results in function approximation setting for MFGs (Huang et al., 2023) or MFC (Pasztor et al., 2021) with similar Lipschitz assumptions, our result does not suffer the exponential term $(1+L_T)^H$ (see Remark F.4 for more explanation).

Additional Remarks on $\log|\mathcal{M}|$ Although low log-covering number of function class is regarded as a standard assumption in many MARL works (Cui et al., 2023; Wang et al., 2023), we would like to take the tabular MFGs as an example and provide remarks about the magnitude of $\log|\mathcal{M}|$. Under Assump. B, with appropriate discretization, the ε-cover for all possible transition functions could be $\Omega(\exp(SAH\mathcal{N}_\varepsilon(\Delta(\mathcal{S}))))$, where $\mathcal{N}_\varepsilon(\Delta(\mathcal{S}))$ denotes the covering number of density space $\Delta(\mathcal{S})$ and we omit L_T, L_r. As a result, in the worst case, $\log|\mathcal{M}| = \Omega(\mathcal{N}_\varepsilon(\Delta(\mathcal{S})))$, which could be exponential in SA. Nonetheless, there are many examples, such that, even in the worst case, $\log|\mathcal{M}|$ is acceptable. For instance, if the model class is parameterized by some $\theta \in \Theta$ (e.g. Neural Networks)

Algorithm 1: `MEBP`: **M**odel **E**limination via **B**ridge **P**olicy

Input: Model Class $\mathcal{M}$; Parameters $\varepsilon_0, \widetilde{\varepsilon}, \bar{\varepsilon}, \delta$.
Initialize: $\mathcal{M}^1 \leftarrow \mathcal{M}$, $\delta_0 \leftarrow \frac{\delta}{\log_2 |\mathcal{M}|+1}$
for $k = 1, 2, ...$ **do**
  $\pi^k \leftarrow \arg\min_\pi |\mathcal{B}_\pi^{\varepsilon_0}(M_{\text{Ctr}}^\pi; \mathcal{M}^k)|$;
  if $|\mathcal{B}_{\pi^k}^{\varepsilon_0}(M_{Ctr}^{\pi^k}; \mathcal{M}^k)| \leq \frac{|\mathcal{M}^k|}{2}$ **then** $\mathcal{M}^{k+1} \leftarrow \texttt{ModelElim}(\pi^k, \mathcal{M}^k, \widetilde{\varepsilon}, \delta_0)$;
  else
    $\pi_{\text{Br}}^{\text{NE},k} \leftarrow \texttt{BridgePolicy}(\mathcal{M}^k, \bar{\varepsilon})$; $\mathcal{M}^{k+1} \leftarrow \texttt{ModelElim}(\pi_{\text{Br}}^{\text{NE},k}, \mathcal{M}^k, \widetilde{\varepsilon}, \delta_0)$;
    Randomly pick $\widetilde{M}^k$ from $\mathcal{M}^{k+1}$; $\mathcal{E}_{\widetilde{M}^k}^{\text{NE}}(\pi_{\text{Br}}^{\text{NE},k}) \leftarrow \max_\pi \Delta_{\widetilde{M}^k}(\pi, \pi_{\text{Br}}^{\text{NE},k})$
    if $\mathcal{E}_{\widetilde{M}^k}^{NE}(\pi_{Br}^{NE,k}) \leq \frac{3\varepsilon}{4}$ **then return** $\pi_{Br}^{NE,k}$;
  end
  if $|\mathcal{M}^k| = 1$ **then** Return the NE of the model ;
end

and take the concatenation of $[s, a, \mu] \in \mathbb{R}^{\dim(\mathcal{S})+\dim(\mathcal{A})+S}$ as inputs, where we use $\dim(\cdot)$ to denote the dimension of a given set. Then $\log|\mathcal{M}| = \widetilde{O}(\dim(\Theta))$, which could just scale with $\widetilde{O}(\text{Poly}(\dim(\mathcal{S}), \dim(\mathcal{A}), S)\})$. As another example, when the transition function only depends on some sufficient statistics of density instead of the exact density, (e.g. $\mathbb{P}(\cdot|s, a, \mu) = \mathbb{P}(\cdot|s, a, \mathbb{E}_{\widetilde{s}\sim\mu}[\widetilde{s}])$), we may have $\log|\mathcal{M}| = \widetilde{O}(\text{Poly}(S, A, H))$ in the worst case. Note that, for the single-agent RL, the largest log-covering number of models is also bounded by $\widetilde{O}(\text{Poly}(S, A, H))$ (folklore).

Exponential Separation between MFGs and MFC Different from MFGs, in Mean-Field Control (MFC) setting, agents cooperate to find an optimal policy to maximize the total return. Previous work (Huang et al., 2023) suggests that both MFC and MFGs can be solved via a unified MLE framework with similar sample complexity upper bounds. One natural question is: *whether learning MFC can also be as sample-efficient as single-agent RL*?

We provide a negative answer to this question. In Thm. F.9, we show that even in tabular setting, there exists a hard instance such that learning MFC requires $\Omega(\exp(|\mathcal{S}|))$ samples. This suggests an exponential separation between learning MFC and MFGs from information-theoretical perspective. Intuitively, for MFC, in the worst case, the agent should explore the entire $\mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$ space to identify the policy that achieves the maximal return. In contrast, as we will explain in Lem. 4.3, in MFGs setting, the learner does not have to explore the entire state-action-density space; instead, finding a "locally-aligned equilibrium policy" is enough.

4.2. Algorithm Design and Proof Sketch

4.2.1. ModelElim: The Model Elimination Step

`ModelElim` can be arbitrary single-agent model elimination procedures. Here we provide an example in Alg. 2.

The basic idea of Alg. 2 is to eliminate models not aligned with M^* conditioning on the given reference policy π. In each iteration, we first find a tuple $(\widetilde{\pi}^t, M^t, M'^t)$ resulting in the maximal discrepancy $\Delta_{\max}^t$. As long as $\Delta_{\max}^t > \widetilde{\varepsilon}$, we collect samples and remove models with low likelihood. With high probability, on the one hand, M^* will never be ruled out under Assump. A; on the other hand, the growth of $\sum_t \Delta_{\max}^t$ is controlled by P-MBED. As a result, the algorithm will terminate eventually and return a model class only including those M with small $d(M, M^*|\pi)$. We summarize the result in the theorem below.

Theorem 4.2. *[Informal version of Thm. F.3] Given any reference policy $\pi, \widetilde{\varepsilon}, \delta \in (0, 1)$, if $M^* \in \bar{\mathcal{M}}$, by choosing $T = \widetilde{O}(\frac{H^4 \dim_{\text{PE}}(\mathcal{M}, \varepsilon')}{\widetilde{\varepsilon}^2})$ with $\varepsilon' = O(\frac{\widetilde{\varepsilon}}{H^2(1+L_T)^H})$, w.p. $1-\delta$, Alg. 2 terminates at some $T_0 \leq T$, and return $\bar{\mathcal{M}}^{T_0}$ s.t. (i) $M^* \in \bar{\mathcal{M}}^{T_0}$ (ii) $\forall M \in \bar{\mathcal{M}}^{T_0}$, $d(M^*, M|\pi) \leq \widetilde{\varepsilon}$.*

We claim Alg. 2 is a single-agent model elimination subroutine, because from Line 8-11, we can see that Alg. 2 only eliminates those $M \in \mathcal{M}^k$ s.t. $\mathbb{P}_{M,h}(\cdot|\cdot, \cdot, \mu_{M,h}^\pi)$ distinct from $\mathbb{P}_{M^*,h}(\cdot|\cdot, \cdot, \mu_{M^*,h}^\pi)$ under some adversarial policy $\widetilde{\pi}^t$ or π itself. Here the density part of $\mathbb{P}_{M,h}$ is fixed by μ_M^π, so during the elimination, all the transitions reduce to single-agent functions only depending on states and actions.

Beyond P-MBED Notably, although we focus on P-MBED in this paper, one may consider other complexity measures generalized from single-agent RL setting (Foster et al., 2021; Sun et al., 2019) and our analysis can be extended correspondingly. That's because as long as `ModelElim` satisfies the (i) and (ii) in Thm. 4.2, it can be arbitrary and does not affect the function of other components in Alg. 1.

4.2.2. Fast Elimination with Bridge Policy

We seek to construct reference policies that allow to eliminate models as efficient as possible, more specifically, to

Algorithm 2: Model Elimination given a Policy

Input: Reference Policy π; $\bar{\mathcal{M}}$; $\widetilde{\varepsilon}, \delta$
Initialize: $\bar{\mathcal{M}}^1 \leftarrow \bar{\mathcal{M}}$; $\mathcal{Z}^0 \leftarrow \{\}$; Set T by Thm. F.3;
for $t = 1, 2, ..., T$ **do**
  $\widetilde{\pi}^t \leftarrow \arg\max_{\widetilde{\pi}} \max_{M, M' \in \bar{\mathcal{M}}^t} \mathbb{E}_{\widetilde{\pi}, M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h}) - \mathbb{P}_{M',h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M',h})\|_1]$.
  $\Delta^t_{\max} \leftarrow$ the maximal value achieved above.
  if $\Delta^t_{\max} \leq \widetilde{\varepsilon}$ **then return** $\bar{\mathcal{M}}^t$;
  else
    for $h = 1, 2..., H$ **do**
      Query Sampling Oracle (Def. 2.1) with (π, π); collect the data at step h: $z^t_h := \{(s^t_h, a^t_h, s'^t_{h+1})\}$.
      Query Sampling Oracle (Def. 2.1) with $(\widetilde{\pi}^t, \pi)$; collect the data at step h: $\widetilde{z}^t_h := \{(s^t_h, a^t_h, s'^t_{h+1})\}$.
      $\mathcal{Z}^t \leftarrow \mathcal{Z}^{t-1} \cup z^t_h \cup \widetilde{z}^t_h$.
    end
    $\forall M \in \bar{\mathcal{M}}^t, l^{\pi}_{\text{MLE}}(M; \mathcal{Z}^t) := \sum_{i \in [t], h \in [H]} \log \mathbb{P}_{M,h}(s'^i_{h+1}|s^i_h, a^i_h, \mu^{\pi}_{M,h}) + \log \mathbb{P}_{M,h}(\widetilde{s}'^i_{h+1}|\widetilde{s}^i_h, \widetilde{a}^i_h, \mu^{\pi}_{M,h})$.
    $\bar{\mathcal{M}}^{t+1} \leftarrow \{M \in \bar{\mathcal{M}}^t | \, l^{\pi}_{\text{MLE}}(M; \mathcal{Z}^t) \geq \max_{\widetilde{M}} l^{\pi}_{\text{MLE}}(\widetilde{M}; \mathcal{Z}^t) - \log \frac{HT|\mathcal{M}|}{\delta}\}$.
  end
end
return $\bar{\mathcal{M}}^T$.

halve the model candidates every iteration until finding the NE. We first consider the simple case, where the models are "scattered" and easy to be distinguished: there exists a policy π^k, such that, no more than $|\mathcal{M}^k|/2$ models are around its CM (resp. `If`-branch, Line 5 in Alg. 1). In this case, after running `ModelElim` with π^k, $\mathcal{M}^{k+1}$ only contains those models locating at the neighborhood of M^* conditioning π^k, which implies $|\mathcal{M}^{k+1}| \leq |\mathcal{B}^{\varepsilon_0}_{\pi^k}(M^{\pi^k}_{\text{Ctr}}; \mathcal{M}^k)| \leq |\mathcal{M}^k|/2$.

The challenging scenario is that, for any policy, the corresponding CM is surrounded by over a half of models (resp. `Else`-branch, Line 6 in Alg. 1). In that case, unstrategically selecting reference policies leads to inefficient model elimination. We present a subtle choice of reference policy, called *Bridge Policy*, which can be constructed by Alg. 3. Before diving into the details of our constructions, we first explain the key insights behind it. Our first insight is summarized in the lemma below.

Lemma 4.3. *[Implication of Local Alignment] Given any $M, \widetilde{M}$ with transition $\mathbb{P}_M$ and $\mathbb{P}_{\widetilde{M}}$, denote $\widehat{\pi}^{NE}_M$ to be an ε_1-approximate NE of M, suppose $d(M, \widetilde{M}|\widehat{\pi}^{NE}_M) \leq \varepsilon_2$, then $\widehat{\pi}^{NE}_M$ is also an $O(\varepsilon_1 + \varepsilon_2)$-approximate NE of $\widetilde{M}$.*

Lem. 4.3 states that, if two models M and $\widetilde{M}$ align with each other conditioning on the NE of one of them, then they approximately share that NE. Therefore, in the `Else`-branch, after calling `ModelElim` with $\pi^{\text{NE},k}_{\text{Br}}$ as the reference policy, if the NE-Gap $\mathcal{E}^{\text{NE}}_{\widetilde{M}^k}(\pi^{\text{NE},k}_{\text{Br}})$ is small for some randomly selected $\widetilde{M}^k \in \mathcal{M}^{k+1}$ (resp. Line 9), we can claim $\pi^{\text{NE},k}_{\text{Br}}$ is an approximate NE of M^* by Lem. 4.3.

However, the remaining challenge is that, if $\mathcal{E}^{\text{NE}}_{\widetilde{M}^k}(\pi^{\text{NE},k}_{\text{Br}})$ is large, we cannot conclude anything about it. Hence, $\pi^{\text{NE},k}_{\text{Br}}$ should be chosen in a strategic way, so that in this case, we can guarantee the elimination is efficient, i.e. $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$. Our second key insight to overcome this challenge is summarized in Thm. 4.4, which indicates that the existence of a "Bridge Policy" that coincides with the NE of its corresponding CM.

Theorem 4.4. *[Bridge Policy] If the `Else`-branch in Line 6 in Alg. 1 is activated, running Alg. 3 returns a bridge policy $\pi^{NE,k}_{Br}$, such that, $\pi^{NE,k}_{Br}$ is an approximate NE of $M^{\pi^{NE,k}_{Br}}_{Ctr}$.*

Before we explain how to prove Thm. 4.4, we first check the implication of this result. Based on Thm. 4.4, if $\mathcal{E}^{\text{NE}}_{\widetilde{M}^k}(\pi^{\text{NE},k}_{\text{Br}}) > \frac{3\varepsilon}{4}$ in Line 9, by Lem. 4.3, we know $d(M^{\pi^{\text{NE},k}_{\text{Br}}}_{\text{Ctr}}, M^*|\pi^{\text{NE},k}_{\text{Br}})$ cannot be small. Therefore, with appropriate hyperparameter choices, we can assert that all models in the neighborhood $\mathcal{B}^{\varepsilon_0}_{\pi^{\text{NE},k}_{\text{Br}}}(M^{\pi^{\text{NE},k}_{\text{Br}}}_{\text{Ctr}}, \mathcal{M}^k)$ should have been eliminated, implying $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$.

Combining the discussions above, we know our Alg. 1 guarantees to either return an approximate NE, or at least halve the model sets. We summarize to the following theorem, which paves the way to our main theorem Thm. 4.1.

Theorem 4.5. *In Alg. 1, by choosing $\varepsilon_0 = \frac{\varepsilon}{8(1+L_r H)(H+4)}$, $\widetilde{\varepsilon} = \frac{\varepsilon_0}{6}$, and choosing $\bar{\varepsilon}$ according to Thm. F.5, w.p. $1-\delta$, (1) if the `If-Branch` in Line 5 is activated: we have $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$; (2) otherwise, in the `Else-Branch` in Line 6: either we return the $\pi^{NE,k}_{Br}$ which is an ε-approximate NE for M^*; or the algorithm continues with $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$.*

Algorithm 3: Bridge Policy Construction

1 **Input**: MF-MDP model class $\mathcal{M}$; $\varepsilon_0, \bar{\varepsilon}$;
2 Convert $\mathcal{M}$ to a PAM class $\ddot{\mathcal{M}}$ via Eq. (1).
3 Construct $\bar{\varepsilon}$-cover of the policy space Π w.r.t. $d_{\infty,1}$, denoted as $\Pi_{\bar{\varepsilon}}$ (see Def. F.1).
4 **for** $\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}$ **do** Find the Central Model $\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\widetilde{\pi}; \ddot{\mathcal{M}}) \leftarrow \arg\max_{\ddot{M} \in \ddot{\mathcal{M}}} |\mathcal{B}^{\varepsilon_0}_{\pi}(\ddot{M}; \ddot{\mathcal{M}})|$;
5 Construct the new PAM $\ddot{M}_{\text{Br}}$ s.t. for any s_h, a_h, π,

$$\ddot{\mathbb{P}}_{\text{Br},h}(\cdot|s_h, a_h, \pi) := \frac{\sum_{\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}} [2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ \ddot{\mathbb{P}}_{\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\widetilde{\pi}; \ddot{\mathcal{M}}),h}(\cdot|s_h, a_h, \widetilde{\pi})}{\sum_{\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}} [2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+},$$

$$\ddot{r}_{\text{Br},h}(s_h, a_h, \pi) := \frac{\sum_{\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}} [2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ \ddot{r}_{\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\widetilde{\pi}; \ddot{\mathcal{M}}),h}(s_h, a_h, \widetilde{\pi})}{\sum_{\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}} [2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+}, \qquad \text{where } [x]^+ := \max\{0, x\}$$

6 Compute the NE of $\ddot{M}_{\text{Br}}$: $\pi^{\text{NE}}_{\text{Br}} \leftarrow \arg\min_{\pi} \max_{\widetilde{\pi}} \ddot{J}_{\ddot{M}_{\text{Br}}}(\widetilde{\pi}; \pi) - \ddot{J}_{\ddot{M}_{\text{Br}}}(\pi; \pi)$.
7 **return** $\pi^{\text{NE}}_{\text{Br}}$.

Proof Sketch of Thm. 4.4 An informal way to interpret the existence of such bridge policy in Thm. 4.4 is to consider a mapping $\mathcal{T}$ from an arbitrary $\pi \in \Pi$ to the NE of its CM M^{π}_{Ctr}. Then, Thm. 4.4 states that $\mathcal{T}$ has an approximate fixed point $\pi^{\text{NE}}_{\text{Br}} \approx \mathcal{T}(\pi^{\text{NE}}_{\text{Br}})$. However, given that it's hard to evaluate the continuity of $\mathcal{T}$ and moreover, $\mathcal{T}$ can be a one-to-many mapping if multiple NEs exist, we prove the existence of such $\pi^{\text{NE}}_{\text{Br}}$ by the non-trivial construction in Alg. 3. We leave the connection between our proofs and the fixed-point theorems as an open problem.

Before explaining our construction in Alg. 3, we first introduce a new notion called *"Policy-Aware Model"* (abbr. PAM) denoted by $\ddot{M}$. The main motivation for introducing PAM is that we want to focus on the policy space, because the feasible densities $\{\mu^{\pi}_{M^*,h}\}_{\pi \in \Pi}$ may not cover the entire density space $\Delta(\mathcal{S}_h)$, and it is not easy to characterize. We defer to Appx. E.1 for the formal definition of PAM and also new notations in Alg. 3 (e.g. $\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\cdot, \cdot)$ denotes Central Model, $\ddot{J}$ denotes the total return), and only summarize the main idea here to save space. Briefly speaking, a PAM $\ddot{M} := \{\mathcal{S}, \mathcal{A}, \mu_1, H, \ddot{\mathbb{P}}, \ddot{r}\}$ is an MDP whose transition $\ddot{\mathbb{P}} : \mathcal{S} \times \mathcal{A} \times \Pi \to \Delta(\mathcal{S})$ and reward functions $\ddot{r} : \mathcal{S} \times \mathcal{A} \times \Pi \to [0, \frac{1}{H}]$ depend on state, action and a "reference policy". PAM can be regarded as a higher-level abstraction of MF-MDP (i.e. MF-MDP $\subset$ PAM), where we replace the dependence on μ^{π}_{M} in MF-MDP by π. We can convert a MF-MDP M to a PAM $\ddot{M}$ sharing the same $\mathcal{S}, \mathcal{A}, \mu_1, H$ by assigning the following for any $h \in [H]$ with $\mu^{\pi}_{M,1} = \mu_1, \forall \pi \in \Pi$:

$$\begin{aligned} \ddot{\mathbb{P}}_{\ddot{M},h}(\cdot|\cdot, \cdot, \pi) :=& \mathbb{P}_{M,h}(\cdot|\cdot, \cdot, \mu^{\pi}_{M,h}) \\ \ddot{r}_{\ddot{M},h}(\cdot, \cdot, \pi) :=& r_h(\cdot, \cdot, \mu^{\pi}_{M,h}), \ \mu^{\pi}_{M,h+1} \leftarrow \Gamma^{\pi}_{M,h}(\mu^{\pi}_{M,h}). \end{aligned} \tag{1}$$

In Alg. 3, we first convert each MF-MDP to its PAM version. Then, we find an $\bar{\varepsilon}$-cover of the policy space w.r.t. $d_{\infty,1}$, denoted by $\Pi_{\bar{\varepsilon}}$, and construct the "Bridge PAM" $\ddot{M}_{\text{Br}}$ by interpolating among CMs w.r.t. $\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}$. Here the weights $[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+$ is chosen carefully for the following considerations:

- **(I)** since $\Pi_{\bar{\varepsilon}}$ is an $\bar{\varepsilon}$-cover, for any $\pi \in \Pi$, the denominator $\sum_{\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}} [2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+$ is always larger than $\bar{\varepsilon}$, which implies both $\ddot{\mathbb{P}}_{\text{Br},h}$ and $\ddot{r}_{\text{Br},h}$ are well-defined and continuous in π. The continuity is important since it implies that $\ddot{M}_{\text{Br}}$ has at least one NE (Def. E.1), denoted as $\pi^{\text{NE}}_{\text{Br}}$;
- **(II)** $[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+$ decays to zero if π largely disagrees with $\widetilde{\pi}$, so $\ddot{\mathbb{P}}_{\text{Br}}(\cdot|\cdot, \cdot, \pi)$ is only determined by CMs of those $\widetilde{\pi}$ close to π.

Next, we discuss what we can conclude from the above two points. Based on the triggering condition in Line 6 in Alg. 1, for any $\pi, \widetilde{\pi}$, the neighbors of M^{π}_{Ctr} and $M^{\widetilde{\pi}}_{\text{Ctr}}$ share at least one common model M_{share}. By using M_{share} as a bridge, we have $\|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}}}(\cdot|\cdot, \cdot, \pi) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}}}(\cdot|\cdot, \cdot, \widetilde{\pi})\|_1 = O(d_{\infty,1}(\pi, \widetilde{\pi}))$. Combining with **(II)**, we know $\forall \pi$, $\|\ddot{\mathbb{P}}_{\text{Br}}(\cdot|\cdot, \cdot, \pi) - \ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}}}(\cdot|\cdot, \cdot, \pi)\|_1 = O(\bar{\varepsilon})$, which implies $\ddot{\mathbb{P}}_{\text{Br}}(\cdot|\cdot, \cdot, \pi^{\text{NE}}_{\text{Br}}) \approx \mathbb{P}_{\ddot{M}^{\pi^{\text{NE}}_{\text{Br}}}_{\text{Ctr}}}(\cdot|\cdot, \cdot, \pi^{\text{NE}}_{\text{Br}})$ if $\bar{\varepsilon}$ is small enough. By the definition of NE in PAM, the conversion rules in Eq. (1) and Lem. 4.3, we can conclude that $\pi^{\text{NE}}_{\text{Br}}$ is an approximate NE of $M^{\pi^{\text{NE}}_{\text{Br}}}_{\text{Ctr}}$, and finish the proof of Thm. 4.4.

5. Learning in Multi-Type MFGs

In this section, we extend our results to the more general Multi-Type MFGs setting[3], allowing to address heterogeneous agents.

Reduction to Lifted MFGs with Constrained Policy Our key observation is that, a MT-MFG $\boldsymbol{M}$ to can be lift to a new MF-MDP $M_{\text{MFG}} := \{\mathcal{S}_{\text{MFG}}, \mathcal{A}_{\text{MFG}}, \mu_1, H, \mathbb{P}_{\text{MFG}}, r_{\text{MFG}}\}$ by augmenting the original states and actions with the type index. The new state and action spaces are given by: $\mathcal{S}_{\text{MFG}} := \bigcup_{w \in [W]} \{\mathcal{S}^w \times \{w\}\}$ and $\mathcal{A}_{\text{MFG}} := \bigcup_{w \in [W]} \{\mathcal{A}^w \times \{w\}\}$.

[3]The existence of NE in MT-MFGs can be found in Thm. E.12.

We defer the detailed description for the conversion process and the definition of initial state distribution, transition and reward functions in M_{MFG} to Appx. G.2.

For policies in M_{MFG}, we only consider $\Pi^\dagger := \{\pi | \forall w \in [W], \pi(a^w \circ w | s^w \circ w) = \pi^w(a^w|s^w),\ \pi^w \in \Pi^w\}$, including all policies which only take actions with the same type as the states. Similar to the NE defined in full policy space Π, we can define the "constrained NE" when agents are constrained to only take policies in the subset $\Pi^\dagger$. More concretely, we call $\widehat{\pi}^{\text{NE}}_{\text{Cstr}} \in \Pi^\dagger$ the ε-approximate Constrained Nash Equilibrium if $\forall \pi \in \Pi^\dagger,\ J_{M_{\text{MFG}}}(\pi, \widehat{\pi}^{\text{NE}}_{\text{Cstr}}) \leq J_{M_{\text{MFG}}}(\widehat{\pi}^{\text{NE}}_{\text{Cstr}}, \widehat{\pi}^{\text{NE}}_{\text{Cstr}}) + \varepsilon$. The following property reveals the connection between constrained NE in M_{MFG} and the NE in the original multi-type model $\boldsymbol{M}$.

Proposition 5.1. *Given a MT-MFG $\boldsymbol{M}$ and its lifted MFG M_{MFG}, we have: (1) an ε-constrained NE $\widehat{\pi}^{NE}_{Cstr} \in \Pi^\dagger$ for M_{MFG} is a $(W\varepsilon)$-NE in $\boldsymbol{M}$; (2) an ε-NE $\widehat{\boldsymbol{\pi}}^{NE}$ in $\boldsymbol{M}$ is an ε-constrained NE for M_{MFG}.*

The above result not only implies the existence of constrained NE in M_{MFG} given the existence of NE in $\boldsymbol{M}$ by letting $\varepsilon \to 0$, but also suggests one can solve NE of MT-MFG by solving the constrained NE in its lifted MFG. The second point is very important since the constrained NE can be solved via almost the same procedures in Sec. 4, as long as we constrain the policy space to $\Pi^\dagger$. We defer algorithm details to Appx. G.5, and summarize our main result in the following theorem.

Theorem 5.2. *[Informal version of Thm. G.8] Under Assump. C and D, there exists an algorithm (Alg. 4), s.t. w.p. $1-\delta$, it returns an ε-NE of $\boldsymbol{M}^*$ after consuming at most*

$$\widetilde{O}\left(\frac{W^2 H^7}{\varepsilon^2}(1+\boldsymbol{L}_r H)^2 \dim_{\text{MTPE}}(\boldsymbol{\mathcal{M}}, \varepsilon')\right)$$

trajectories, where $\varepsilon' = O\left(\frac{\varepsilon}{WH^3(1+\boldsymbol{L}_r H)(1+\boldsymbol{L}_T)^H}\right)$.

In Appx. H, we investigate a practical multi-agent system called N-player Multi-Type Symmetric Anonymous Games (MT-SAGs) generalized from SAGs. We establish approximation error between MT-MFGs and MT-SAGs. Our results reveal a larger class of Multi-Agent systems where NE can be solved in a sample-efficient way.

6. A Heuristic Algorithm with Improved Computational Efficiency

Although Alg. 1 is sample-efficient, it requires exponential computation. In this section, we aim to design a heuristic algorithm[4] sharing the main insights as Alg. 1 while more computationally tractable. For the lack of space, we defer the concrete algoirthm (Alg. 7), the experiment setting and evaluation results (Fig. 2) to Appx. J. In this section, we just highlight the algorithm design.

Highlights of Algorithm Design We assume a NE Oracle is available, such that given a known MFG model, the Oracle can return its NE. We argue that such oracle can be easily implemented if the model is smooth enough or the monotonicity condition is satisfied (Guo et al., 2019; Perolat et al., 2021). Besides, in our experiments, we observe that repeatedly mixing the policy with its best response can converge to a good solution. Given such oracle, Alg. 7 only involves $|\mathcal{M}|$ calls of NE oracle, and $\text{Poly}(|\mathcal{M}|, |\mathcal{S}|, |\mathcal{A}|, H)$ arithmetic operations in computing model difference or likelihood, which avoids exponential computation in Alg. 1.

For the algorithm design, Alg. 7 follows the same if-else structure as Alg. 2, but we improve the computational efficiency in two aspects. Firstly, we avoid procedures optimizing over the entire policy class, including Line 4 in Alg. 1 and Line 4 in Alg. 2. Instead, we only search over the NE policies of model candidates, which can be computed by calling the NE Oracle $|\mathcal{M}|$ times at the beginning. As long as the models in $\mathcal{M}$ are diverse enough, we can expect their NEs to be reasonable representatives for Π in distinguishing models. Secondly, we replace the $\pi^{\text{NE},k}_{\text{Br}}$ in Alg. 1 with the NE of the model $M^k \leftarrow \arg\max_{M \in \mathcal{M}^k} |\mathcal{B}^{\varepsilon_0}_{\pi^{\text{NE}}_M}(M, \mathcal{M}^k)|$, and do not have to solve the NE of the complicated bridge model in Alg. 3. We claim that this modification still aligns with Alg. 3 in principle. Note that the main intuition behind Alg. 3 is that, when Line 6 in Alg. 1 is activated, the reference policy used for elimination should be a policy π_{ref}, such that, π_{ref} collapses with the NE of the model with the maximal number of neighbors conditioning on π_{ref}.

7. Conclusion

In this paper, we reveal that learning MFGs can be as sample-efficient as single-agent RL under mild assumptions, and the sample complexity of RL in MFGs can be characterized by a novel complexity measure called Partial Model-Based Eluder Dimension (P-MBED). Besides, we extend our algorithms to the more general Multi-Type MFGs setting. Lastly, we contribute an empirical algorithm with improved computational efficiency.

As for the future, one interesting direction is to study the sample complexity when only value function approximations are available. Besides, while our focus is the sample efficiency in this paper, it would be valuable to identify general conditions, under which computationally efficient algorithms exist. Lastly, our results underscore the power of mean-field approximation, and it would be worthwhile to investigate other generalizations of the MFGs setting, in order to deepen our understanding on the sample efficiency of learning NE in other MARL systems.

[4]The code is available at https://github.com/jiaweihhuang/Heuristic_MEBP.

Impact Statement

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

Acknowledgements

This research was supported by Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343, SNSF starting grant, SNSF grant agreement 51NF40 180545, and by the European Research Council (ERC) under the European Union's Horizon grant 815943.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Anahtarci, B., Kariksiz, C. D., and Saldi, N. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

Cardaliaguet, P. and Lehalle, C.-A. Mean field game of controls and an application to trade crowding. *Mathematics and Financial Economics*, 12:335–363, 2018.

Chen, Z., Li, C. J., Yuan, A., Gu, Q., and Jordan, M. I. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022a.

Chen, Z., Zhou, D., and Gu, Q. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *International Conference on Algorithmic Learning Theory*, pp. 227–261. PMLR, 2022b.

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

Cui, Q., Zhang, K., and Du, S. S. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. *arXiv preprint arXiv:2302.03673*, 2023.

Daskalakis, C., Golowich, N., and Zhang, K. The complexity of markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4180–4234. PMLR, 2023.

Djehiche, B., Tcheukam, A., and Tembine, H. Mean-field-type games in engineering. *arXiv preprint arXiv:1605.03281*, 2016.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Elie, R., Perolat, J., Laurière, M., Geist, M., and Pietquin, O. 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.

Foster, D., Foster, D. J., Golowich, N., and Rakhlin, A. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2678–2792. PMLR, 2023.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Ghosh, A. and Aggarwal, V. Model free reinforcement learning algorithm for stationary mean field equilibrium for multiple types of agents. *arXiv preprint arXiv:2012.15377*, 2020.

Gomes, D. A. and Pimentel, E. A. Economic models and mean-field games theory.

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

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

Huang, J., Chen, J., Zhao, L., Qin, T., Jiang, N., and Liu, T.-Y. Towards deployment-efficient reinforcement learning: Lower bound and optimality. *arXiv preprint arXiv:2202.06450*, 2022.

Huang, J., Yardim, B., and He, N. On the statistical efficiency of mean field reinforcement learning with general function approximation, 2023.

Huang, M., Malhamé, R. P., and Caines, P. E. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. 2006.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021b.

Lasry, J.-M. and Lions, P.-L. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

Laurière, M., Perrin, S., Geist, M., and Pietquin, O. Learning mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2022.

Levy, O., Cassel, A., Cohen, A., and Mansour, Y. Eluder-based regret for stochastic contextual mdps, 2022.

Mao, W., Qiu, H., Wang, C., Franke, H., Kalbarczyk, Z., Iyer, R., and Basar, T. A mean-field game approach to cloud resource management with function approximation. In *Advances in Neural Information Processing Systems*, 2022.

Mishra, R. K., Vasal, D., and Vishwanath, S. Model-free reinforcement learning for non-stationary mean field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 1032–1037. IEEE, 2020.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Moon, J. and Başar, T. Linear quadratic mean field stackelberg differential games. *Automatica*, 97:200–213, 2018.

Ni, C., Song, Y., Zhang, X., Jin, C., and Wang, M. Representation learning for general-sum low-rank markov games. *arXiv preprint arXiv:2210.16976*, 2022.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

Pasztor, B., Bogunovic, I., and Krause, A. Efficient model-based multi-agent mean-field reinforcement learning. *arXiv preprint arXiv:2107.04050*, 2021.

Perolat, J., Perrin, S., Elie, R., Laurière, M., Piliouras, G., Geist, M., Tuyls, K., and Pietquin, O. Scaling up mean field games with online mirror descent. *arXiv preprint arXiv:2103.00623*, 2021.

Perrin, S., Pérolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Subramanian, S. G., Poupart, P., Taylor, M. E., and Hegde, N. Multi type mean field reinforcement learning. *arXiv preprint arXiv:2002.02513*, 2020.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

uz Zaman, M. A., Miehling, E., and Başar, T. Reinforcement learning for non-stationary discrete-time linear–quadratic mean-field games in multiple populations. *Dynamic Games and Applications*, 13(1):118–164, 2023.

Vasal, D. and Berry, R. Master equation for discrete-time stackelberg mean field games with a single leader. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 5529–5535. IEEE, 2022.

Wang, Y., Liu, Q., Bai, Y., and Jin, C. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. *arXiv preprint arXiv:2302.06606*, 2023.

Xie, Q., Yang, Z., Wang, Z., and Minca, A. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pp. 11436–11447. PMLR, 2021.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Yardim, B., Cayci, S., Geist, M., and He, N. Policy mirror ascent for efficient and independent learning in mean field games. *arXiv preprint arXiv:2212.14449*, 2022.

Yardim, B., Goldman, A., and He, N. When is mean-field reinforcement learning tractable and relevant? *arXiv preprint arXiv:2402.05757*, 2024.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Zhang, F., Tan, V. Y., Wang, Z., and Yang, Z. Learning regularized monotone graphon mean-field games. *arXiv preprint arXiv:2310.08089*, 2023.

Zhang, K., Yang, Z., and Basar, T. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2022.

Outline of the Appendix

- Appx. A: Frequently used notations.
- Appx. B: Additional related works.
- Appx. C: Informal Discussions about how to extend our results to unknown reward function setting.
- Appx. D: Missing details and proofs related to P-MBED.
- Appx. E: Missing details related to Single/Multi-Type Policy Aware Models (PAM).
- Appx. F: Proofs for lemma and theorems related to learning MFGs in Sec. 4.
- Appx. G: Proofs for lemma and theorems related to learning MultiType MFGs in Sec. 5.
- Appx. H: Introduction to Multi-Type Symmetric Anonymous Games (MT-SAGs) and approximation error between MT-MFGs and MT-SAGs.
- Appx. I: Some basic lemma useful in our proofs.
- Appx. J: Experiment details and results.

A. Frequently Used Notations

Notation	Explanation
M	Mean-Field MDP
$\mathcal{M}$	Model class for (single-type) Mean-Field MDP
π	Non-stationary policy for (single-type) Mean-Field MDP
Q_M, V_M, J_M	Value functions for (single-type) MF-MDP
$\dim_{\mathrm{PE}}(\mathcal{M}, \varepsilon)$	Type-I Partial-MBED, Def. 3.3
$\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}, \varepsilon)$	Type-II Partial-MBED, Def. D.1
$d(\cdot, \cdot \vert \pi)$	Conditional distance given a reference policy π
$\mathcal{B}^{\varepsilon_0}_{\pi}(M, \mathcal{M})$	ε_0-neighborhood of M in $\mathcal{M}$ conditioning on π
$M^{\varepsilon_0}_{\mathrm{Ctr}}(\pi, \mathcal{M})$	The "Central Model"
$d_{\infty,1}(\pi, \pi')$	Policy distance
$\boldsymbol{M}$	Multi-Type Mean-Field MDP
$\boldsymbol{\mathcal{M}}$	Model class for multi-type Mean-Field MDP
$\boldsymbol{\pi}$	Non-stationary policy for multi-type Mean-Field MDP
$Q_{\boldsymbol{M}}, V_{\boldsymbol{M}}, J_{\boldsymbol{M}}$	Value functions for multi-type MF-MDP
$\dim_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}}, \varepsilon)$	Type-I Multi-Type Partial-MBED, Def. D.10
$\dim^{\mathrm{II}}_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}}, \varepsilon)$	Type-II Multi-Type Partial-MBED, Def. D.10
$\ddot{M}/\ddot{\boldsymbol{M}}$	Policy-aware model (single-type/multi-type)
$\ddot{\mathcal{M}}/\ddot{\boldsymbol{\mathcal{M}}}$	Model class for the policy-aware model (single-type/multi-type)
$\ddot{Q}_{\ddot{M}}, \ddot{V}_{\ddot{M}}, \ddot{J}_{\ddot{M}}/\ddot{Q}_{\ddot{\boldsymbol{M}}}, \ddot{V}_{\ddot{\boldsymbol{M}}}, \ddot{J}_{\ddot{\boldsymbol{M}}}$	Value functions for the policy-aware model (single-type/multi-type)

B. Additional Related Works

Single-Agent/Multi-Agent RL with General Function Approximation For the single-agent RL with function approximation setting, besides the literature we mentioned in the main text, there are multiple other insightful works (Ayoub et al., 2020; Chen et al., 2022a; Huang et al., 2022; Modi et al., 2021; Uehara et al., 2021; Xie et al., 2022; Zanette et al., 2020; Zhong et al., 2022).

As for the multi-agent setting, sample complexity of Markov Games has been extensively studied in both tabular (Bai et al., 2020; Chen et al., 2022b; Jin et al., 2021b; Zhang et al., 2019; 2021) and function approximation setting (Cui et al., 2023; Foster et al., 2023; Huang et al., 2021; Ni et al., 2022; Wang et al., 2023). These papers study a general MARL setting with individually distinct agents, which is quite different from our MFG or MT-MFG. Besides, many of them study the decentralized training setting, which requires much less communication cost than our centralized setting. However, because of the difficulty in learning NE in general Markov Games setting, most of them focus on the convergence to weaker notions of equilibrium instead, e.g. the Correlated Equilibrium or the Coarse Correlated Equilibria, and those results in function approximation setting (Cui et al., 2023; Wang et al., 2023) may still depend on the number of agents, although in polynomial. In contrast, although we specify in mean-field approximation setting, we can have more ambitious goals on solving Nash Equilibrium, and our sample complexity bounds are totally independent w.r.t. the number of agents. Moreover, we also reveal some cases when learning (MT-)MFG can be as sample-efficient as single-agent RL by investigating the Partial Model-Based Eluder Dimension.

C. Extension to the Setting when the Reward Function is Unknown

We remark that our current results extend to the unknown reward setting. Below we elucidate the key modifications needed for this extension.

Firstly, for the problem setup, we instead assume a model class $\mathcal{M}$ available, where each element $M := (r_M, \mathbb{P}_M) \in \mathcal{M}$ corresponds to a (reward, transition) tuple. The definition of the P-MBED can be amended by incorporating both reward and transition differences in Def. 3.3.

Secondly, for the algorithm design:

- For Algorithm 1, we redefine the model distance the definition $d^{\widetilde{\pi}}(M, \widetilde{M}|\pi)$ (introduced at the beginning of Sec. 4.1) to include the expectation of distances in both reward and transition functions:

  $$d^{\widetilde{\pi}}(M, \widetilde{M}|\pi) := \mathbb{E}_{\widetilde{\pi}, M(\pi)}[\sum_{h=1}^{H} \|r_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h}) - r_{\widetilde{M},h}(\cdot|\cdot,\cdot,\mu^{\pi}_{\widetilde{M},h})\|_1 + \|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h}) - \mathbb{P}_{\widetilde{M},h}(\cdot|\cdot,\cdot,\mu^{\pi}_{\widetilde{M},h})\|_1].$$

  The definition of ε_0-neighborhood and "Central Model" will adjust correspondingly.

- For Algorithm 2: we should augment the reward difference into the right-hand side of Line 4, integrate reward into the dataset in Lines 9 and 11, and include the likelihood of reward functions in Line 13.

- For Algorithm 3, the construction of bridge policy will follow the new definition of model distance $d^{\widetilde{\pi}}(M, \widetilde{M}|\pi)$.

Finally, for the analysis, based on the modified algorithms, under realizability assumption, we can extend Lemma D.7 and prove that the accumulative estimation errors of reward and transition are controlled by P-MBED. The current analysis can be seamlessly generalized to establish sample complexity upper bounds depending on the P-MBED of reward and transition function classes.

D. Missing Details about Partial Model-Based Eluder Dimension

D.1. Alternative Notions of Partial MBED in MFGs

In this section, we introduce a choice of ν different from the one in Def. 3.2, which also leads to a valid P-MBED. In the following, we introduce another choice of ν, which results in a different P-MBED. We will call it "Type II" P-MBED to distinguish the one in Def. 3.3.

Definition D.1 (Type II Partial MBED)**.** Given a model class $\mathcal{M}$, define the mapping $\nu^{\pi}_{M^*,h} : \mathcal{M} \to \Delta(\mathcal{S}_h)$ such that $\forall M \in \mathcal{M}$, $\nu^{\pi}_{M^*,h}(M) := \mu^{\pi}_{M^*,h}$, then the type II P-MBED of $\mathcal{M}$ is defined by:

$$\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}, \varepsilon) := \max_{\pi} \max_{h\in[H]} \dim_{\mathrm{E}|\nu^{\pi}_{M^*,h}}(\mathcal{M}, \varepsilon).$$

We want to highlight here that each of the two types P-MBED has advantages over the other. As we will see in the Thm. F.7, if we use $\dim^{\mathrm{II}}_{\mathrm{PE}}$ to derive the sample complexity upper bound, we have to suffer the exponential term of $(1 + L_T)^H$. On the other hand, in the following proposition, we can see $\dim^{\mathrm{II}}_{\mathrm{PE}}$ is directly comparable with MBED (Huang et al., 2023) ($\alpha = 1$ case), while we can not have the similar guarantee for $\dim_{\mathrm{PE}}$.

Proposition D.2 (Low MBED $\subset$ Low Type II P-MBED)**.** $\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}, \varepsilon) \leq \dim_{\mathrm{E}}(\mathcal{M}, \varepsilon)$.

D.2. Proofs Related to P-MBED in MFGs Setting

Proposition 3.4. *(Tabular Setting) For any $\mathcal{M}$ and $\varepsilon > 0$, $\dim_{\mathrm{PE}}(\mathcal{M}, \varepsilon) \leq |\mathcal{S}||\mathcal{A}|$, while there exists a concrete example of $\mathcal{M}$ such that $\dim_{\mathrm{E}}(\mathcal{M}, \varepsilon) = \Omega(\exp(|\mathcal{S}|))$.*

Proof. When the density is fixed, any MF-MDP reduces to a single-agent MDP. For any single-agent MDP, there are at most $|\mathcal{S}||\mathcal{A}|$ different (s_h, a_h) pairs, for any h. Therefore, the P-MBED can be upper bounded by $|\mathcal{S}||\mathcal{A}|$.

In contrast, for MBED in (Huang et al., 2023), we consider the model class constructed in Thm. F.9. Consider the sequence $\{(s^1_h, a^1_h, \mu^i)\}_{i\in[n]}$ with $\mu^i \in \mathcal{U}_{\zeta=\lfloor \frac{L_T}{5\varepsilon} \rfloor}$ for all $i \in [n]$, but $\mu^i \neq \mu^j$ if $i \neq j$. For any $i \in [n-1]$, there exists two models $\mathbb{P}_{\mu^i}$ and $\mathbb{P}_{\mu^{i+1}}$, such that,

$$\sum_{t=1}^{i-1} \|\mathbb{P}_{\mu^i}(\cdot|s^1_h, a^1_h, \mu^t) - \mathbb{P}_{\mu^{i+1}}(\cdot|s^1_h, a^1_h, \mu^t)\|^2_1 = 0$$

but

$$\|\mathbb{P}_{\mu^i}(\cdot|s^1_h, a^1_h, \mu^i) - \mathbb{P}_{\mu^{i+1}}(\cdot|s^1_h, a^1_h, \mu^i)\|_1 = 4\varepsilon.$$

Note that $|\mathcal{U}_{\zeta=\lfloor \frac{L_T}{5\varepsilon} \rfloor}| = O((\frac{L_T}{S\varepsilon})^{S-1})$, by choosing $\varepsilon \leq \frac{L_T}{2S}$, we have $\dim_{\mathrm{E}}(\mathcal{M}_h, \varepsilon) = \Omega(\exp(S))$. □

Similarly, we can show the type II P-MBED in tabular setting can also be upper bounded by $|\mathcal{S}||\mathcal{A}|$, because there are at most $|\mathcal{S}||\mathcal{A}|$ different state-action tuples.

Proposition D.3 (Type II P-MBED in the Tabular Setting)**.** $\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}, \varepsilon) \leq |\mathcal{S}||\mathcal{A}|$.

Next, we study the linear setting. Given a mapping $f : \mathcal{S} \to \mathbb{R}^d$, we use $\mathrm{Rank}([f(x)]_{x\in\mathcal{X}})$ to denote the rank of matrix concatenated by $[f(x)]_{x\in\mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|\times d}$.

Proposition D.4 (Linear Setting; Formal version of Prop. 3.5)**.** *Consider the Low-Rank MF-MDP with known feature $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \to \mathbb{R}^d$ satisfying $\|\phi\| \leq C_\phi$, and unknown next state feature $\psi : \mathcal{S} \to \mathbb{R}^d$. Given a next state feature function class Ψ satisfying $\forall \psi \in \Psi$, $\forall s' \in \mathcal{S}$, $\forall g : \mathcal{S} \to \{-1, 1\}$, $\|\sum_{s'} \psi(s')g(s')\|_2 \leq C_\Psi$, consider the following model class:*

$$\mathcal{M}_\Psi := \{\mathbb{P}_\psi | \mathbb{P}_\psi(\cdot|s, a, \mu) := \phi(s, a, \mu)^\top \psi(s'); \forall s, a, \mu,\ \mathbb{P}_\psi(\cdot|s, a, \mu) \in \Delta(\mathcal{S}); \psi \in \Psi\},$$

we have $\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}_\Psi, \varepsilon) = \widetilde{O}(\max_{\pi,h} \mathrm{Rank}([\phi_h(s_h, a_h, \mu^{\pi}_{M^*,h})]_{s_h\in\mathcal{S}, a_h\in\mathcal{A}}))$.

Moreover, if $\phi(s, a, \mu)$ has decomposition: $\phi(s, a, \mu)^\top = \phi(s, a)^\top G(\mu)$ with $\phi(\cdot, \cdot) \in \mathbb{R}^{\widetilde{d}}$ and $G(\cdot) \in \mathbb{R}^{\widetilde{d}\times d}$, we have $\dim_{\mathrm{PE}}(\mathcal{M}_\Psi, \varepsilon) = \widetilde{O}(\widetilde{d})$ and $\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}_\Psi, \varepsilon) = \widetilde{O}(\min\{\widetilde{d}, d\})$.

Remark D.5. *As we can see, the P-MBED is related to the "activated dimension" of features after partially fixing the density, which can be much lower than its MBED $\approx d$. Moreover, when the feature is decomposable, the dimension of state-action feature will also serve as an upper bound.*

Proof. **Proof for Type-II P-MBED (Def. D.1)** In the following, we first consider a fixed policy π and h. To simplify the notation, we denote $\Phi := [\phi(s_h, a_h, \mu^{\pi}_{M^*,h})]_{s\in\mathcal{S}, a\in\mathcal{A}} \in \mathbb{R}^{d\times|S||A|}$ to be the matrix concatenated by vectors $\phi(\cdot,\cdot,\mu^{\pi}_{M^*,h})$, and denote $d_{\text{active}} := \text{Rank}(\Phi)$ to be its rank. We use $U := [u_1, u_2, ..., u_{d_{\text{active}}}] \in \mathbb{R}^{d\times d_{\text{active}}}$ to denote a normalized orthogonal basis in $\text{Span}(\Phi) = \text{Span}(U)$ satisfying $\|u_i\|_2 = 1$ for all $i \in [d_{\text{active}}]$ and $u_i^\top u_j = 0$ for any $i \neq j$. Easy to verify that for any s_h, a_h, the following equation

$$U\phi_{\text{active}}(s_h, a_h, \mu^{\pi}_{M^*,h}) = \phi(s_h, a_h, \mu^{\pi}_{M^*,h}).$$

has a solution satisfying:

$$\|\phi_{\text{active}}(s_h, a_h, \mu^{\pi}_{M^*,h})\|_2 = \|U^\top U\phi_{\text{active}}(s_h, a_h, \mu^{\pi}_{M^*,h})\|_2 = \|U^\top\phi(s_h, a_h, \mu^{\pi}_{M^*,h})\|_2 \le \|\phi(s_h, a_h, \mu^{\pi}_{M^*,h})\|_2 \le C_\phi.$$

Given a fixed policy π, $h \in [H]$, suppose $(s_h^1, a_h^1), ..., (s_h^n, a_h^n)$ is a partially ε-independent sequence w.r.t. $\mathcal{M}_\Psi$ and $\nu^{\pi}_{M^*,h}$ defined in D.1. Then for each $i \in [n]$, there should exists $\psi^i, \widetilde{\psi}^i \in \Psi$, such that:

$$\varepsilon^2 \ge \sum_{t=1}^{i-1} \|\mathbb{P}_{\psi^i}(\cdot|s_h^t, a_h^t, \mu^{\pi}_{M^*,h}), \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s_h^t, a_h^t, \mu^{\pi}_{M^*,h})\|_1^2.$$

and

$$\begin{aligned}
\varepsilon^2 \le& \|\mathbb{P}_{\psi^i}(\cdot|s_h^i, a_h^i, \mu^{\pi}_{M^*,h}) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|_1^2 \\
=& \Big(\phi_{\text{active}}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})^\top U^\top \sum_{s'\in\mathcal{S}} (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h}, s')\Big)^2 \\
\le& \|\phi_{\text{active}}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|^2_{(\Lambda_h^i)^{-1}} \|U^\top \sum_{s'\in\mathcal{S}} (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h}, s')\|^2_{\Lambda_h^i}.
\end{aligned}$$

where we define:

$$\Lambda_h^i := \lambda I + \sum_{t=1}^{i-1} \phi_{\text{active}}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\phi_{\text{active}}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})^\top \in \mathbb{R}^{d_{\text{active}}\times d_{\text{active}}};$$

$$g_{\psi^i,\widetilde{\psi}^i}(s_h, a_h, \mu, s') := \begin{cases} 1, & \text{if } \phi_{\text{active}}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})^\top U^\top(\psi^i(s') - \widetilde{\psi}^i(s')) \ge 0; \\ -1, & \text{otherwise.} \end{cases}$$

For simplicity, we use $v_{\psi,\widetilde{\psi}}(s_h, a_h, \mu) := U^\top \sum_{s'}(\psi(s') - \widetilde{\psi}(s'))g_{\psi,\widetilde{\psi}}(s_h, a_h, \mu, s')$ as a shortnote. Therefore, for each i,

$$\begin{aligned}
&\|v_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|^2_{\Lambda_h^i} \\
=& \lambda\|v_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|^2 + \sum_{t=1}^{i-1}\Big(\phi_{\text{active}}(s_h^t, a_h^t, \mu^{\pi}_{M^*,h})^\top v_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\Big)^2 \\
=& \lambda\|v_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|^2 + \sum_{t=1}^{i-1}\Big(\phi_{\text{active}}(s_h^t, a_h^t, \mu^{\pi}_{M^*,h})^\top U^\top \sum_{s'}(\psi^i(s) - \widetilde{\psi}^i(s'))g_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h}, s')\Big)^2 \\
\le& 4\lambda C_\Psi^2 + \sum_{t=1}^{i-1}\|\mathbb{P}_{\psi^i}(\cdot|s_h^t, a_h^t, \mu^{\pi}_{M^*,h}) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s_h^t, a_h^t, \mu^{\pi}_{M^*,h})\|_1^2 \\
\le& 4\lambda C_\Psi^2 + \varepsilon^2.
\end{aligned}$$

By choosing $\lambda = \varepsilon^2/4C_\Psi^2$, we have:

$$\|\phi_{\text{active}}(s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|^2_{(\Lambda_h^i)^{-1}} \ge \frac{\varepsilon^2}{4\lambda C_\Psi^2 + \varepsilon^2} = \frac{1}{2}.$$

On the one hand,

$$\det \Lambda_h^{n+1} = \det(\Lambda_h^n + \phi_{\text{active}}(s_h^n, a_h^n, \mu^\pi_{M^*,h})\phi_{\text{active}}(s_h^n, a_h^n, \mu^\pi_{M^*,h})^\top) = (1 + \|\phi_{\text{active}}(s_h^n, a_h^n, \mu^\pi_{M^*,h})\|^2_{(\Lambda_h^n)^{-1}}) \cdot \det \Lambda_h^n$$
$$\geq \frac{3}{2} \det \Lambda_h^n \geq (\frac{3}{2})^n \det \Lambda_h^1 = \lambda_h^{d_{\text{active}}} (\frac{3}{2})^n.$$

Therefore,

$$\lambda_h^{d_{\text{active}}} (\frac{3}{2})^n \leq \det \Lambda_h^{n+1} \leq (\frac{\text{Tr}(\Lambda_h^n)}{d_{\text{active}}})^{d_{\text{active}}} \leq (\lambda + \frac{nC_\phi^2}{d_{\text{active}}})^{d_{\text{active}}}.$$

which implies $n = O(d_{\text{active}} \log(1 + \frac{d_{\text{active}} C_\phi C_\Psi}{\varepsilon}))$.

Finally, if we take the maximum over all policy π, we have

$$\dim^{\text{II}}_{\text{PE}}(\mathcal{M}_\Psi, \varepsilon) = \widetilde{O}(\max_{\pi,h} \text{Rank}([\phi_h(s_h, a_h, \mu^\pi_{M^*,h})]_{s_h \in \mathcal{S}, a_h \in \mathcal{A}})).$$

When $\phi(s_h, a_h, \mu)$ can be decomposed to $\phi(s_h, a_h)^\top G(\mu)$ for some $\phi(s_h, a_h) \in \mathbb{R}^{\widetilde{d}}$, easy to verify that for any π, the corresponding $d_{\text{active}} \leq \widetilde{d}$. By combining with Prop. D.2, we can conclude $\dim^{\text{II}}_{\text{PE}}(\mathcal{M}, \varepsilon) = \widetilde{O}(\min\{d, \widetilde{d}\})$.

Proofs for P-MBED (Def. 3.3) As for the first type of P-MBED, we only study the decomposable feature setting. Given a fixed policy π, $h \in [H]$, suppose $(s_h^1, a_h^1), ..., (s_h^n, a_h^n)$ is a partially ε-independent sequence w.r.t. $\mathcal{M}_\Psi$ and the mapping ν_h^π defined in Def. 3.3, then for each $i \in [n]$, there should exists $\psi^i, \widetilde{\psi}^i \in \Psi$, such that:

$$\varepsilon^2 \geq \sum_{t=1}^{i-1} \|\mathbb{P}_{\psi^i}(\cdot|s_h^t, a_h^t, \mu^\pi_{\psi^i,h}), \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s_h^t, a_h^t, \mu^\pi_{\widetilde{\psi}^i,h})\|_1^2.$$

and

$$\begin{aligned}
\varepsilon^2 \leq & \|\mathbb{P}_{\psi^i}(\cdot|s_h^i, a_h^i, \mu^\pi_{\psi^i,h}) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s_h^i, a_h^i, \mu^\pi_{\widetilde{\psi}^i,h})\|_1^2 \\
= & \Big(\phi(s_h^i, a_h^i)^\top \sum_{s' \in \mathcal{S}} (G(\mu^\pi_{\psi^i,h})\psi^i(s') - G(\mu^\pi_{\widetilde{\psi}^i,h})\widetilde{\psi}^i(s'))\widetilde{g}_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, s')\Big)^2 \\
\leq & \|\phi(s_h^i, a_h^i)\|^2_{(\Lambda_h^i)^{-1}} \| \sum_{s' \in \mathcal{S}} (G(\mu^\pi_{\psi^i,h})\psi^i(s') - G(\mu^\pi_{\widetilde{\psi}^i,h})\widetilde{\psi}^i(s'))\widetilde{g}_{\psi^i,\widetilde{\psi}^i}(s_h^i, a_h^i, s')\|^2_{\Lambda_h^i}.
\end{aligned}$$

where we define:

$$\Lambda_h^i := \lambda I + \sum_{t=1}^{i-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top \in \mathbb{R}^{\widetilde{d} \times \widetilde{d}};$$

$$g_{\psi^i,\widetilde{\psi}^i}(s_h, a_h, s') := \begin{cases} 1, & \text{if } \phi(s_h^i, a_h^i)^\top (G(\mu^\pi_{\psi^i,h})\psi^i(s') - G(\mu^\pi_{\widetilde{\psi}^i,h})\widetilde{\psi}^i(s')) \geq 0; \\ -1, & \text{otherwise.} \end{cases}$$

The rest analysis is similiar to the non-decomposable setting above. As a result, we can show:

$$n = O(\widetilde{d} \log(1 + \frac{\widetilde{d} C_\phi C_\Psi}{\varepsilon})).$$

This holds for any π, which finishes the proof. □

Remark D.6. *Following similar analyses as Prop. D.4 and Prop. B.6 and Prop. B.7 in (Huang et al., 2023), we can compute the P-MBED for kernel MF-MDP and generalized linear function classes. All we need to do is to replace d_{eff} or d in (Huang et al., 2023) with the corresponding dimensions conditioning on the adversarial densities.*

Lemma D.7. *Under Def. 3.3 and Def. D.1, consider a fixed π and an arbitrary $h \in [H]$, Suppose we have a sequence $\{\mathbb{P}_{M^k,h}\}_{k=1}^K \in \mathcal{F}$ and $\{(s_h^k, a_h^k)\}_{k=1}^K \subset \mathcal{S} \times \mathcal{A}$,*

- *if for all* $k \in [K]$, $\sum_{i=1}^{k-1} \|\mathbb{P}_{M^k,h}(\cdot|s_h^i, a_h^i, \mu^{\pi}_{M^k,h}) - \mathbb{P}_{M^*,h}(\cdot|s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|_1^2 \leq \beta$, *then for any* $\varepsilon > 0$, *we have* $\sum_{k=1}^{K} \|\mathbb{P}_{M^k,h}(\cdot|s_h^k, a_h^k, \mu^{\pi}_{M^k,h}) - \mathbb{P}_{M^*,h}(\cdot|s_h^k, a_h^k, \mu^{\pi}_{M^*,h})\|_1 = O(\sqrt{\beta K \dim_{\mathrm{PE}}(\mathcal{M}, \varepsilon)} + K\varepsilon)$,
- *if for all* $k \in [K]$, $\sum_{i=1}^{k-1} \|\mathbb{P}_{M^k,h}(\cdot|s_h^i, a_h^i, \mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^*,h}(\cdot|s_h^i, a_h^i, \mu^{\pi}_{M^*,h})\|_1^2 \leq \beta$, *then for any* $\varepsilon > 0$, *we have* $\sum_{k=1}^{K} \|\mathbb{P}_{M^k,h}(\cdot|s_h^k, a_h^k, \mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^*,h}(\cdot|s_h^k, a_h^k, \mu^{\pi}_{M^*,h})\|_1 = O(\sqrt{\beta K \dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}, \varepsilon)} + K\varepsilon)$.

Proof. Let's consider a single-agent model-class $\mathcal{P} \subset \{P | P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$. We first introduce the notion of independent state-action sequences given a single-agent model class.

Definition D.8 (ε-Independent state action pairs)**.** Given single-agent model $\mathcal{P}$ and a data sequence $\{(s_h^i, a_h^i)\}_{i=1}^n \subset \mathcal{S}_h \times \mathcal{A}_h$, we say (s_h, a_h) is ε-independent of $\{(s_h^i, a_h^i)\}_{i=1}^n$ w.r.t. $\mathcal{P}$ if there exists $\mathbb{P}, \widetilde{\mathbb{P}} \in \mathcal{M}$ such that $\sum_{i=1}^n \|\mathbb{P}_h(\cdot|s_h^i, a_h^i) - \widetilde{\mathbb{P}}_h(\cdot|s_h^i, a_h^i)\|_1^2 \leq \varepsilon^2$ but $\|\mathbb{P}_{M,h}(\cdot|s_h, a_h) - \widetilde{\mathbb{P}}_h(\cdot|s_h, a_h)\|_1 > \varepsilon$. We call $\{(s_h^i, a_h^i)\}_{i=1}^n$ an ε-independent sequence w.r.t. $\mathcal{P}$ (at step h) if for any $i \in [n]$, (s_h^i, a_h^i, μ_h^i) is ε-independent w.r.t. $\{(s_h^t, a_h^t, \mu_h^t)\}_{t=1}^{i-1}$.

We use $\dim_{\mathrm{E}}(\mathcal{P}, \varepsilon)$ to denote the maximal length of ε-independent sequence $\{(s^i, a^i)\}_{i \in [n]}$ for single-agent model class $\mathcal{P}$. Since single-agent RL is a special case of MF-MDP where the transition is independent w.r.t. density. As implied by Lem. 4.4 in (Huang et al., 2023) when $\alpha = 1$, suppose there is a sequence $\{\mathbb{P}^k\}_{k \in [K]} \subset \mathcal{P}$ and a sequence of states and actions $\{(s_h^k, a_h^k)\}_{k \in K}$, such that:

$$\sum_{i=1}^{k-1} \|\mathbb{P}^k(\cdot|s_h^i, a_h^i) - \mathbb{P}^*(\cdot|s_h^i, a_h^i)\|_1^2 \leq \beta$$

where $\mathbb{P}^* \in \mathcal{P}$ is some fixed function, then for any $\varepsilon > 0$,

$$\sum_{k=1}^{K} \|\mathbb{P}^k(\cdot|s_h^k, a_h^k) - \mathbb{P}^*(\cdot|s_h^k, a_h^k)\|_1 \leq O(\sqrt{\beta K \dim_{\mathrm{E}}(\mathcal{P}, \varepsilon)} + K\varepsilon).$$

By choosing $\mathcal{P} := \{\mathbb{P}_{M,h} | \mathbb{P}_{M,h}(\cdot|\cdot,\cdot) \leftarrow \mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h}), M \in \mathcal{M}\}$ with $\mathbb{P}^* := \mathbb{P}_{M^*,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^*,h})$ and combining the definition in Def. 3.3, we can finish the proof of the first statement.

By choosing $\mathcal{P} := \{\mathbb{P}_{M,h} | \mathbb{P}_{M,h}(\cdot|\cdot,\cdot) \leftarrow \mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^*,h}), M \in \mathcal{M}\}$ with $\mathbb{P}^* := \mathbb{P}_{M^*,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^*,h})$ and combining the definition in Def. D.1, we can finish the proof of the second statement. □

D.3. Partial MBED for Model Classes in Multi-Type MFGs Setting

Definition D.9 (Partial ε-Independence for Multi-Type Mean-Field Model Classes)**.** Given a multi-type model class $\boldsymbol{\mathcal{M}}$, consider a $w \in [W]$ and a mapping $\nu_h^w : \boldsymbol{\mathcal{M}} \to \Delta(\mathcal{S}_h^1) \times ... \times \Delta(\mathcal{S}_h^W)$, and a sequence of data $\{(s_h^{w,i}, a_h^{w,i})\}_{i=1}^n \subset \mathcal{S}_h^w \times \mathcal{A}_h^w$, we say (s_h^w, a_h^w) is partially ε-independent on $\{(s_h^{w,i}, a_h^{w,i})\}_{i=1}^n$ w.r.t. $\boldsymbol{\mathcal{M}}$ and ν_h^w at step h, if there exists $\boldsymbol{M}, \widetilde{\boldsymbol{M}} \in \boldsymbol{\mathcal{M}}$, s.t. $\sum_{i=1}^n \|\mathbb{P}^w_{\boldsymbol{M},h}(\cdot|s_h^{w,i}, a_h^{w,i}, \nu_h^w(\boldsymbol{M})) - \mathbb{P}^w_{\widetilde{\boldsymbol{M}},h}(\cdot|s_h^{w,i}, a_h^{w,i}, \nu_h^w(\widetilde{\boldsymbol{M}}))\|_1^2 \leq \varepsilon^2$ but $\|\mathbb{P}^w_{\boldsymbol{M},h}(\cdot|s_h^w, a_h^w, \nu_h^w(\boldsymbol{M})) - \mathbb{P}^w_{\widetilde{\boldsymbol{M}},h}(\cdot|s_h^w, a_h^w, \nu_h^w(\widetilde{\boldsymbol{M}}))\|_1 > \varepsilon$.

Besides, we call $\{(s_h^{w,i}, a_h^{w,i})\}_{i=1}^n$ is a partially ε-independent sequence w.r.t. $\boldsymbol{\mathcal{M}}$ and ν_h^w if for any $i \in [n]$, $(s_h^{w,i}, a_h^{w,i})$ is partially ε-independent on $\{(s_h^{w,t}, a_h^{w,t})\}_{t=1}^{i-1}$. In the following, we use $\dim_{\mathrm{E}|\nu_h^w}(\boldsymbol{\mathcal{M}}, \varepsilon)$ to denote the length of the longest partially ε-independent sequence w.r.t. $\boldsymbol{\mathcal{M}}$ and ν_h^w for type w (at step h).

Definition D.10. Given a model class $\boldsymbol{\mathcal{M}}$ and an arbitrary w, we define the mapping $\nu_h^{w,\boldsymbol{\pi}} : \boldsymbol{\mathcal{M}} \to \Delta(\mathcal{S}^1) \times ...\Delta(\mathcal{S}^W)$ s.t. $\nu_h^{w,\boldsymbol{\pi}}(\boldsymbol{M}) = \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h}$, and the mapping $\nu^{w,\boldsymbol{\pi}}_{\boldsymbol{M}^*,h} : \boldsymbol{\mathcal{M}} \to \Delta(\mathcal{S}^1) \times ...\Delta(\mathcal{S}^W)$ s.t. $\nu^{w,\boldsymbol{\pi}}_{\boldsymbol{M}^*,h}(\boldsymbol{M}) = \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M}^*,h}$. Then, the Multi-Type P-MBEDs are defined by:

- Type I MT-P-MBED: $\dim_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}}, \varepsilon) := \sum_{w \in [W]} \max_{h \in [H]} \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \dim_{\mathrm{E}|\nu_h^{w,\boldsymbol{\pi}}}(\boldsymbol{\mathcal{M}}, \varepsilon)$;
- Type II MT-P-MBED: $\dim^{\mathrm{II}}_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}}, \varepsilon) := \sum_{w \in [W]} \max_{h \in [H]} \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}} \dim_{\mathrm{E}|\nu^{w,\boldsymbol{\pi}}_{\boldsymbol{M}^*,h}}(\boldsymbol{\mathcal{M}}, \varepsilon)$.

Proposition D.11 (Tabular Multi-Type MF-MDP)**.**

$$\max\{\dim_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}},\varepsilon),\dim^{\mathrm{II}}_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}},\varepsilon)\}\leq\sum_{w\in[W]}|\mathcal{S}^w||\mathcal{A}^w|.$$

Proof. By definition, for each w, for any fixed h and $\boldsymbol{\pi}$, the longest partially independent state-action sequence would have length $|\mathcal{S}^w||\mathcal{A}^w|$. □

Proposition D.12 (Linear Multi-Type MF-MDP)**.** *Consider the Low-Rank Multi-Type MF-MDP with known feature $\phi^w:\mathcal{S}^w\times\mathcal{A}^w\times\Delta(\mathcal{S}^1)\times...\Delta(\mathcal{S}^W)\to\mathbb{R}^{d^w}$ satisfying $\|\phi^w\|\leq C_\phi$ for any $w\in[W]$, and unknown next state feature $\psi^w:\mathcal{S}^w\to\mathbb{R}^d$. Given a next state feature function class $\Psi^1,...,\Psi^W$ satisfying $\forall\psi^w\in\Psi^w$, $\forall s'^w\in\mathcal{S}^w$, $\forall g:\mathcal{S}\to\{-1,1\}$, $\|\sum_{s'^w}\psi(s'^w)g(s'^w)\|_2\leq C_\Psi$, define the model class:*

$$\mathcal{M}_{\Psi^w}:=\{\mathbb{P}_{\psi^w}|\forall s^w,a^w,\boldsymbol{\mu},\ \mathbb{P}^w_\psi(\cdot|s^w,a^w,\boldsymbol{\mu}):=\phi^w(s^w,a^w,\boldsymbol{\mu})^\top\psi^w(s'^w);\ \mathbb{P}^w_\psi(\cdot|s^w,a^w,\boldsymbol{\mu})\in\Delta(\mathcal{S}^w);\psi^w\in\Psi^w\},$$

then, we have:

$$\forall w\in[W],\quad\dim^{\mathrm{II}}_{\mathrm{MTPE}}(\mathcal{M}_{\Psi^w},\varepsilon)=O(\sum_{w\in[W]}\max_{\boldsymbol{\pi}\in\boldsymbol{\Pi},h\in[H]}d^w_{active,\boldsymbol{\pi},h}\log(1+\frac{d^w_{active,\boldsymbol{\pi},h}C_\phi C_\Psi}{\varepsilon}))$$

where $d^w_{active,\boldsymbol{\pi},h}:=\mathrm{Rank}([\phi^w_h(s^w_h,a^w_h,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M}^*,h})]_{s^w_h\in\mathcal{S}^w,a^w_h\in\mathcal{A}^w})$.

Moreover, if $\phi^w(s^w,a^w,\boldsymbol{\mu})$ is decomposable, i.e. for any $w\in[W]$, $\phi^w(s^w,a^w,\boldsymbol{\mu})^\top=\phi^w(s^w,a^w)^\top G^w(\boldsymbol{\mu})$ with $\phi^w(\cdot,\cdot)\in\mathbb{R}^{\widetilde{d}^w}$ and $G^w(\cdot)\in\mathbb{R}^{\widetilde{d}^w\times d^w}$, we have $\dim_{\mathrm{MTPE}}(\mathcal{M}_{\Psi^w},\varepsilon)=\widetilde{O}(\sum_{w\in[W]}\widetilde{d}^w)$ and $\dim^{\mathrm{II}}_{\mathrm{MTPE}}(\mathcal{M}_{\Psi^w},\varepsilon)=\widetilde{O}(\sum_{w\in[W]}\min\{\widetilde{d}^w,d^w\})$.

Proof. The proof is a direct generalization of Prop. D.4 by applying the same techniques in the proof of Prop. D.4 for each type $w\in[W]$, so we omit it here. □

D.4. Partial MBED in Constrained Policy Spaces

Next, we define the Constrained Partial MBED extended from Def. 3.3, where the main difference is that we constrain the set of adversarial policies.

Definition D.13 (Constrained Partial MBED in MFRL)**.** Given a (single-type) Mean-Field model class $\mathcal{M}$, and M^* denotes the true model, we consider the same ν^π_h and $\nu^\pi_{M^*,h}$ function defined in Def. 3.3 and Def. D.1, respectively. Then, the constrained P-MBEDs are defined by:

- Type I P-MBED: $\dim^{\mathrm{I}}_{\mathrm{CPE}|\Pi^\dagger}(\mathcal{M},\varepsilon):=\max_{h\in[H]}\max_{\pi\in\Pi^\dagger}\dim_{\mathrm{E}|\nu^\pi_h}(\mathcal{M},\varepsilon)$;
- Type II P-MBED: $\dim^{\mathrm{II}}_{\mathrm{CPE}|\Pi^\dagger}(\mathcal{M},\varepsilon):=\max_{h\in[H]}\max_{\pi\in\Pi^\dagger}\dim_{\mathrm{E}|\nu^\pi_{M^*,h}}(\mathcal{M},\varepsilon)$.

Comparing with P-MBED, the main difference is that in constrained P-MBED the adversarial policies are only chosen from the constrained policy set. Recall the definition of $\Pi^\dagger$ in Sec. 5. Given a $\boldsymbol{\mathcal{M}}$ and a model class $\mathcal{M}_{\mathrm{MFG}}$ converted from $\boldsymbol{\mathcal{M}}$ according to Appx. G.2, we have the following relationship between the P-MBED of $\mathcal{M}_{\mathrm{MFG}}$ constrained on $\Pi^\dagger$ and P-MBED of $\boldsymbol{\mathcal{M}}$.

Proposition D.14. *Given a model class $\boldsymbol{\mathcal{M}}$ and its corresponding lifted MFG class $\mathcal{M}_{MFG}$:*

$$\dim_{\mathrm{CPE}|\Pi^\dagger}(\mathcal{M}_{MFG},\varepsilon)\leq\dim_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}},\varepsilon),\tag{2}$$

$$\dim^{\mathrm{II}}_{\mathrm{CPE}|\Pi^\dagger}(\mathcal{M}_{MFG},\varepsilon)\leq\dim^{\mathrm{II}}_{\mathrm{MTPE}}(\boldsymbol{\mathcal{M}},\varepsilon).\tag{3}$$

Proof. Let's consider a fixed policy $\pi\in\Pi^\dagger$. Note that, π corresponds to a $\boldsymbol{\pi}:=\{\pi^w\}_{w\in[W]}$ with $\pi^w:\mathcal{S}^w\to\Delta(\mathcal{A}^w)$ and $\pi^w(a^w_h|s^w_h)=\pi(a^w_h\circ w|s^w_h\circ w)$. Given any $\varepsilon>0$, and $h\in[H]$, suppose we have a partial ε-independent sequence w.r.t. the mapping ν^π_h (or $\nu^\pi_{M^*,h}$), denoted as $\{(s^{w_i,i}_h\circ w_i,a^{w_i,i}_h\circ w_i)\}_{i\in[n]}$. We divide this sequence according to its group w_i, which we denote as $\{\{s^{w,i_w}_h\circ w,a^{w,i_w}_h\circ w\}_{i_w\in[n_w]}\}_{w\in[W]}$ with $\sum_w n_w=n$. By construction of $\mathcal{M}_{\mathrm{MFG}}$, for any $w\in[W]$,

$\{s_h^{w,i_w}, a_h^{w,i_w}\}_{i_w \in [n_w]}$ is a partial ε-independent sequence w.r.t. function class $\mathcal{M}^w$ and the mapping $\nu_h^{w,\boldsymbol{\pi}}$ (or $\nu_{\boldsymbol{M}^*,h}^{w,\boldsymbol{\pi}}$), which is upper bounded by the Multi-Type P-MBED of model class $\mathcal{M}^w$.

We finish the proof of Eq. (2) by maximizing over $\pi \in \Pi^\dagger$. □

As directly implied by Prop. D.14, Prop. D.11 and Prop. D.12, we can upper bound the constrained P-MBED in some special cases.

E. Details about Single-Type/Multi-Type Policy Aware Models

E.1. (Single-Type) Policy-Aware Model

Concretely, Policy-Aware Model (PAM) is specified by a tuple $\ddot{M} := \{\mathcal{S}, \mathcal{A}, H, \mu_1, \ddot{\mathbb{P}}_{\ddot{M}}, \ddot{r}_{\ddot{M}}\}$, where $\mathcal{S}, \mathcal{A}, H, \mu_1$ are the state space, action space, horizon length, initial state distribution which are the same as the normal MF-MDP setting; $\ddot{\mathbb{P}}_{\ddot{M}} := \{\ddot{\mathbb{P}}_{\ddot{M},1}, ..., \ddot{\mathbb{P}}_{\ddot{M},H}\}$ is the transition function with $\ddot{\mathbb{P}}_{\ddot{M},h} : \mathcal{S}_h \times \mathcal{A}_h \times \Pi \to \Delta(\mathcal{S}_{h+1})$, and $\ddot{r}_{\ddot{M}} := \{\ddot{r}_{\ddot{M},1}, ..., \ddot{r}_{\ddot{M},H}\}$ is the reward function[5] satisfying $\ddot{r}_{\ddot{M},h} : \mathcal{S}_h \times \mathcal{A}_h \times \Pi \to [0, 1/H]$, where recall Π denotes the collection of all Markov policies. Given any reference policy π, we define the value function $\ddot{Q}^{\widetilde{\pi}}_{\ddot{M},h} : \mathcal{S}_h \times \mathcal{A}_h \times \Pi \to \mathbb{R}$ and $\ddot{V}^{\widetilde{\pi}}_{\ddot{M},h} : \mathcal{S}_h \times \Pi \to \mathbb{R}$ regarding $\widetilde{\pi}$ in the following way:

$$\ddot{Q}^{\widetilde{\pi}}_{\ddot{M},h}(\cdot,\cdot,\pi) := \mathbb{E}\Big[\sum_{h'=h}^{H} \ddot{r}_{\ddot{M},h'}(s_{h'}, a_{h'}, \pi)\Big| s_h = \cdot,\ a_h = \cdot,\ \forall h' \geq h,\ s_{h'+1} \sim \ddot{\mathbb{P}}_{\ddot{M},h'}(\cdot|s_{h'}, a_{h'}, \pi), a_{h'+1} \sim \widetilde{\pi}(\cdot|s_{h'+1})\Big],$$

$$\ddot{V}^{\widetilde{\pi}}_{\ddot{M},h}(\cdot,\pi) := \mathbb{E}\Big[\sum_{h'=h}^{H} \ddot{r}_{\ddot{M},h'}(s_{h'}, a_{h'}, \pi)\Big| s_h = \cdot,\ a_h \sim \widetilde{\pi},\ \forall h' \geq h,\ s_{h'+1} \sim \ddot{\mathbb{P}}_{\ddot{M},h'}(\cdot|s_{h'}, a_{h'}, \pi), a_{h'+1} \sim \widetilde{\pi}(\cdot|s_{h'+1})\Big].$$

Similarly, we will denote $\mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\cdot]$ to be the expectation taken over trajectories sampled by executing $\widetilde{\pi}$ in the model $\ddot{M}$, such that the transition and reward functions are fixed by π. Again, we will call π as the "reference policy".

By definition, once the reference policy π is determined, the transition/reward functions reduced to single-agent transition/reward functions, and the value functions are defined in the same way as single-agent RL setting. Besides, we define the total return of $\widetilde{\pi}$ conditioning on the reference policy π as:

$$\ddot{J}_{\ddot{M}}(\widetilde{\pi}; \pi) := \mathbb{E}_{s_1 \sim \mu_1}[\ddot{V}^{\widetilde{\pi}}_{\ddot{M},h}(s_1, \pi)],$$

and define $\Delta_{\ddot{M}}(\widetilde{\pi}, \pi) := \ddot{J}_{\ddot{M}}(\widetilde{\pi}, \pi) - \ddot{J}_{\ddot{M}}(\pi, \pi)$. Similar to MF-MDP, we define the NE in $\ddot{M}$. Intuitively, the NE in $\ddot{M}$ is the policy $\pi^{\text{NE}}_{\ddot{M}}$ that agents do not tend to deviate when $\pi^{\text{NE}}_{\ddot{M}}$ is chosen to be the reference policy.

Definition E.1 (Nash Equilibrium in $\ddot{M}$)**.** Given a model $\ddot{M}$, we call $\pi^{\text{NE}}_{\ddot{M}}$ is a Nash Equilibrium (NE) of $\ddot{M}$, if

$$\forall \widetilde{\pi} \in \Pi, \quad \ddot{J}_{\ddot{M}}(\widetilde{\pi}; \pi^{\text{NE}}_{\ddot{M}}) \leq \ddot{J}_{\ddot{M}}(\pi^{\text{NE}}_{\ddot{M}}; \pi^{\text{NE}}_{\ddot{M}}).$$

Besides, we call $\widehat{\pi}^{\text{NE}}_{\ddot{M}}$ is an ε-approximate NE of $\ddot{M}$, if

$$\forall \widetilde{\pi} \in \Pi, \quad \ddot{J}_{\ddot{M}}(\widetilde{\pi}; \pi^{\text{NE}}_{\ddot{M}}) \leq \ddot{J}_{\ddot{M}}(\pi^{\text{NE}}_{\ddot{M}}; \pi^{\text{NE}}_{\ddot{M}}) + \varepsilon.$$

Similar to the conditional distance $d(M, \widetilde{M}|\pi)$ defined in Sec. 4, we can define the conditional distance for PAM.

$$d(\ddot{M}, \ddot{\widetilde{M}}|\pi) := \max_{\widetilde{\pi}}\ \max\{\mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M},h}(\cdot|\cdot,\cdot,\pi) - \ddot{\mathbb{P}}_{\ddot{\widetilde{M}},h}(\cdot|\cdot,\cdot,\pi)\|_1],$$
$$\mathbb{E}_{\widetilde{\pi},\ddot{\widetilde{M}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M},h}(\cdot|\cdot,\cdot,\pi) - \ddot{\mathbb{P}}_{\ddot{\widetilde{M}},h}(\cdot|\cdot,\cdot,\pi)\|_1]\}$$

Given a PAM model class $\ddot{\mathcal{M}}$ and a model $\ddot{M} \in \ddot{\mathcal{M}}$, for any reference policy π, we define the ε_0-neighborhood of $\ddot{M}$ in $\ddot{\mathcal{M}}$ to be

$$\mathcal{B}^{\varepsilon_0}_{\pi}(\ddot{M}; \ddot{\mathcal{M}}) := \{\ddot{M}' \in \ddot{\mathcal{M}} | d(\ddot{M}, \ddot{M}'|\pi) \leq \varepsilon_0\}.$$

Besides, we define the Central Model $\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\pi; \ddot{\mathcal{M}})$ in $\ddot{\mathcal{M}}$ regarding π to be the model with the largest neighborhood set:

$$\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\pi; \ddot{\mathcal{M}}) \leftarrow \arg\max_{\ddot{M} \in \ddot{\mathcal{M}}} |\mathcal{B}^{\varepsilon_0}_{\pi}(\ddot{M}; \ddot{\mathcal{M}})|.$$

[5] Here we specify the model in the subscription, because for those PAM converted from MF-MDPs, even if they share the reward function in mean-field systems, the reward functions in PAM version can be different because of the difference in transition functions.

E.1.1. EXISTENCE OF NASH EQUILIBRIUM IN $\ddot{M}$

Next, we investigate the existence of NE in $\ddot{M}$. Recall the definition

$$d_{\infty,1}(\pi,\pi') := \max_{h\in[H], s_h\in\mathcal{S}_h} \|\pi_h(\cdot|s_h) - \pi'_h(\cdot|s_h)\|_1. \tag{4}$$

Theorem E.2. *[Existence of Nash Equilibrium in PAM] Given a PAM $\ddot{M}$ with discrete state and action spaces, such that, for any $h \in [H], s_{h+1} \in \mathcal{S}_{h+1}, s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$, both $\ddot{\mathbb{P}}_{\ddot{M},h}(s_{h+1}|s_h,a_h,\pi)$ and $\ddot{r}_{\ddot{M},h}(s_h,a_h,\pi)$ are continuous at π w.r.t. distance $d_{\infty,1}$, then $\ddot{M}$ has at least one NE satisfying Def. E.1.*

Proof. In Prop. E.2, we establish the existence of NE in Multi-Type PAM, and the proof for this theorem is a special case when $W = 1$. □

As a direct result of Thm. E.2 and Lem. E.4, we have the following corollary.

Corollary E.3. *Given a MF-MDP model M satisfying Assump. B, the PAM model $\ddot{M}$ converted from M according to the rules in Eq. (1) has at least one NE.*

E.1.2. USEFUL LEMMA RELATED TO THE PAM CONVERTED FROM MF-MDP

Lemma E.4. *[Lipschitz Continuity of PAM] Given an MF-MDP M satisfying the Lipschitz continuity condition in Assump. B, consider the PAM $\ddot{M}$ converted from M according to Eq. (1), we have $\ddot{M}$ is also Lipschitz continuous that, $\forall h \in [H]$ and any $s_h \in \mathcal{S}, a_h \in \mathcal{A}$,*

$$\|\ddot{\mathbb{P}}_{\ddot{M},h}(\cdot|s_h,a_h,\pi) - \ddot{\mathbb{P}}_{\ddot{M},h}(\cdot|s_h,a_h,\pi')\|_1 \le d_{\infty,1}(\pi,\pi')L_T \sum_{h'=1}^{h}(1+L_T)^{h-h'}$$

$$|\ddot{r}_{\ddot{M},h}(s_h,a_h,\pi) - \ddot{r}_{\ddot{M},h}(s_h,a_h,\pi')| \le d_{\infty,1}(\pi,\pi')L_r \sum_{h'=1}^{h}(1+L_T)^{h-h'}.$$

Proof. This lemma is a special case of Lem. E.13 when $W = 1$. □

E.2. Multi-Type Policy-Aware Model

In this section, we introduce Multi-Type Policy-Aware Model (MT-PAM) extended from PAM. To distinguish with MT-MFG, we use $\ddot{\boldsymbol{M}}$ as notation.

MT-PAM is specified by $\ddot{\boldsymbol{M}} := \{(\mu_1^w, \mathcal{S}^w, \mathcal{A}^w, H, \ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}}}, \ddot{r}^w_{\ddot{\boldsymbol{M}}})_{w\in[W]}\}$[6], where $\mathcal{S}^w, \mathcal{A}^w, H, \mu_1^w$ are defined the same as the Multi-Type MF-MDP setting; $\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}}} := \{\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}\}_{h\in[H]}$ is the transition function with $\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h} : \mathcal{S}_h^w \times \mathcal{A}_h^w \times \mathbf{\Pi} \to \Delta(\mathcal{S}_{h+1})$ and $\ddot{r}_{\ddot{\boldsymbol{M}},h} : \mathcal{S}_h^w \times \mathcal{A}_h^w \times \mathbf{\Pi} \to [0, \frac{1}{H}]$, where recall $\mathbf{\Pi}$ denotes the set of all Markov policies $\boldsymbol{\pi} := \{\pi^w\}_{w\in[W]}$ with $\pi^w \in \Pi^w$.

Given a reference policy $\boldsymbol{\pi} := \{\pi^w\}_{w\in[W]} \in \mathbf{\Pi}$, for any $\widetilde{\boldsymbol{\pi}} := \{\widetilde{\pi}^w\}_{w\in[W]} \in \mathbf{\Pi}$, we define the value function for type w $\ddot{Q}^{w,(\cdot)}_{\ddot{\boldsymbol{M}},h} : \mathcal{S}_h^w \times \mathcal{A}_h^w \times \mathbf{\Pi} \to \mathbb{R}$ and $\ddot{V}^{w,(\cdot)}_{\ddot{\boldsymbol{M}},h} : \mathcal{S}_h^w \times \mathbf{\Pi} \to \mathbb{R}$ in the following way:

$$\ddot{Q}^{w,\widetilde{\boldsymbol{\pi}}}_{\ddot{\boldsymbol{M}},h}(\cdot,\cdot;\boldsymbol{\pi}) := \mathbb{E}\Big[\sum_{h'=h}^{H} \ddot{r}^w_{\ddot{\boldsymbol{M}},h'}(s^w_{h'}, a^w_{h'}, \boldsymbol{\pi})\Big| s^w_h = \cdot,\ a^w_h = \cdot,\ \forall \widetilde{h} \ge h,\ \ s^w_{\widetilde{h}+1} \sim \ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},\widetilde{h}}(\cdot|s^w_{\widetilde{h}}, a^w_{\widetilde{h}}, \boldsymbol{\pi}),\ a^w_{\widetilde{h}+1} \sim \widetilde{\pi}^w_{\widetilde{h}+1}\Big].$$

$$V^{w,\widetilde{\boldsymbol{\pi}}}_{\ddot{\boldsymbol{M}},h}(\cdot;\boldsymbol{\pi}) := \mathbb{E}\Big[\sum_{h'=h}^{H} \ddot{r}^w_{\ddot{\boldsymbol{M}},h'}(s^w_{h'}, a^w_{h'}, \boldsymbol{\pi})\Big| s^w_h = \cdot,\ a^w_h \sim \widetilde{\pi}^w_h,\ \ \forall \widetilde{h} \ge h,\ \ s^w_{\widetilde{h}+1} \sim \ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},\widetilde{h}}(\cdot|s^w_{\widetilde{h}}, a^w_{\widetilde{h}}, \boldsymbol{\pi}),\ a^w_{\widetilde{h}+1} \sim \widetilde{\pi}^w_{\widetilde{h}+1}\Big].$$

Similarly, we will denote $\mathbb{E}_{\widetilde{\boldsymbol{\pi}},\ddot{M}(\boldsymbol{\pi})}[\cdot]$ to be the expectation taken over trajectories sampled by executing $\widetilde{\boldsymbol{\pi}}$ in the model $\ddot{\boldsymbol{M}}$, such that the transition and reward functions are fixed by $\boldsymbol{\pi}$. Again, we will call $\boldsymbol{\pi}$ as the "reference policy".

[6]Here we specify the model in the subscription of the reward function, which will avoid confusion when we consider the PAMs converted from (Multi-Type) MF-MDPs

We denote $\ddot{J}^w_{\ddot{\boldsymbol{M}}}(\widetilde{\boldsymbol{\pi}};\boldsymbol{\pi}) := \mathbb{E}_{s_1^w\sim\mu_1^w}[V^{w,\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{M},1}(s_1^w;\boldsymbol{\pi})]$ to be the expected return of agents in type w in model $\ddot{\boldsymbol{M}}$ by executing $\widetilde{\boldsymbol{\pi}}$ given $\boldsymbol{\pi}$ as the reference policy.

Definition E.5 (Nash Equilibrium in Multi-Type PAM)**.** The Nash Equilibrium policy in Multi-Type PAM is defined to be the policy $\boldsymbol{\pi}^{\text{NE}} := \{\pi^{w,\text{NE}}\}_{w\in[W]}$ satisfying:

$$\forall w\in[W],\ \forall\widetilde{\boldsymbol{\pi}}\in\boldsymbol{\Pi},\quad \ddot{J}^w_{\ddot{\boldsymbol{M}}}(\widetilde{\boldsymbol{\pi}};\boldsymbol{\pi}^{\text{NE}})\le\ddot{J}^w_{\ddot{\boldsymbol{M}}}(\boldsymbol{\pi}^{\text{NE}};\boldsymbol{\pi}^{\text{NE}}). \tag{5}$$

Note that $\ddot{J}^w_{\ddot{\boldsymbol{M}}}(\widetilde{\boldsymbol{\pi}};\boldsymbol{\pi})$ actually only depends on $\boldsymbol{\pi}$ and $\widetilde{\pi}^w$.

E.2.1. Existence of Nash Equilibrium in MT-PAM

We first investigate a stronger notion of NE, which we call the "strict NE".

Definition E.6 (Strict NE)**.** Given a MT-PAM $\ddot{\boldsymbol{M}}$ with transitions and rewards $\{(\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}}},\ddot{r}^w_{\ddot{\boldsymbol{M}}})\}_{w\in[W]}$, the policy $\boldsymbol{\pi}^{\text{NE}}$ is a strict NE of $\ddot{\boldsymbol{M}}$ if and only if the following holds:

$$\forall w\in[W],\quad \pi^w_h(\cdot|s^w_h)\in\operatorname*{arg\,max}_{u\in\Delta(\mathcal{A}^w)}\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h,\cdot,\boldsymbol{\pi})^\top u. \tag{6}$$

Note that this is a stronger notion than the NE defined in Def. E.5, i.e. a strict NE is always a NE. In the following, we will focus on the existence of strict NE.

Lemma E.7 (Strict NE as Fixed Point)**.** *Given a MT-PAM $\ddot{\boldsymbol{M}}$ with transitions and rewards $\{(\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}}},\ddot{r}^w_{\ddot{\boldsymbol{M}}})\}_{w\in[W]}$, the policy $\boldsymbol{\pi}^{SNE}$ is a strict NE of $\ddot{\boldsymbol{M}}$ if and only if the following holds:*

$$\Gamma^{SNE}_{\ddot{\boldsymbol{M}}}(\boldsymbol{\pi}^{SNE})=\boldsymbol{\pi}^{SNE},$$

where

$$\Gamma^{SNE}_{\ddot{\boldsymbol{M}}}(\boldsymbol{\pi}):=\{\widetilde{\boldsymbol{\pi}}:=\{\widetilde{\pi}^w_h\}_{w\in[W],h\in[H]}|\forall w,s^w_h,\ \widetilde{\pi}^w_h(\cdot|s^w_h):=\operatorname*{arg\,max}_{u\in\Delta(\mathcal{A}^w)}\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h,\cdot,\boldsymbol{\pi})^\top u-\|\pi^w_h(\cdot|s^w_h)-u\|_2^2\}.$$

Proof. First of all, suppose $\boldsymbol{\pi}$ is the NE of $\ddot{\boldsymbol{M}}$ according to Def. E.5, by the policy improvement theorem in single-agent RL, we have:

$$\forall w\in[W],\quad \pi^w_h(\cdot|s^w_h)\in\operatorname*{arg\,max}_{u\in\Delta(\mathcal{A}^w)}\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h,\cdot,\boldsymbol{\pi})^\top u,$$

which also implies

$$\forall w\in[W],\ \pi^w_h(\cdot|s^w_h)\in\arg\max_{u\in\Delta(\mathcal{A}^w)}\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h,\cdot,\boldsymbol{\pi})^\top u-\|\pi^w_h(\cdot|s^w_h)-u\|_2^2.$$

Therefore, if $\boldsymbol{\pi}$ is the strict NE of $\ddot{\boldsymbol{M}}$, we have $\Gamma^{\text{SNE}}_{\ddot{\boldsymbol{M}}}(\boldsymbol{\pi})=\boldsymbol{\pi}$.

On the other hand, if $\Gamma^{\text{SNE}}_{\ddot{M}}(\boldsymbol{\pi})=\boldsymbol{\pi}$, it implies:

$$\forall w\in[W],\ \pi^w_h(\cdot|s^w_h)\in\arg\max_{u\in\Delta(\mathcal{A}^w)}\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h,\cdot,\boldsymbol{\pi})^\top u-\|\pi^w_h(\cdot|s^w_h)-u\|_2^2.$$

By the first order optimality condition of the RHS, we should have:

$$\forall w\in[W],\quad \pi^w_h(\cdot|s^w_h)\in\operatorname*{arg\,max}_{u\in\Delta(\mathcal{A}^w)}\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h,\cdot,\boldsymbol{\pi})^\top u,$$

Therefore, $\boldsymbol{\pi}$ is the strict NE of $\ddot{\boldsymbol{M}}$. □

Definition E.8 (Distance measure between policies)**.** Given two policies $\boldsymbol{\pi}:=\{\pi^w_h\}_{h\in[H],w\in[W]}$ and $\widetilde{\boldsymbol{\pi}}:=\{\widetilde{\pi}^w_h\}_{h\in[H],w\in[W]}$, we define:

$$d_{\infty,1}(\boldsymbol{\pi},\widetilde{\boldsymbol{\pi}}):=\max_{w\in[W],h\in[H],s^w_h\in\mathcal{S}^w_h}\|\pi^w_h(\cdot|s^w_h)-\widetilde{\pi}^w_h(\cdot|s^w_h)\|_1.$$

Condition E.9. Given a model function $\ddot{\boldsymbol{M}} := \{(\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}, \ddot{r}^w_{\ddot{\boldsymbol{M}},h})\}_{h\in[H]}$, for any fixed $w \in [W], h \in [H], s^w_{h+1} \in \mathcal{S}^w_{h+1}, s^w_h \in \mathcal{S}^w_h, a^w_h \in \mathcal{A}^w_h$ and any $\boldsymbol{\pi}$, $\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}(s^w_{h+1}|s^w_h, a^w_h, \boldsymbol{\pi})$ and $\ddot{r}^w_{\ddot{\boldsymbol{M}},h}(s^w_h, a^w_h, \boldsymbol{\pi})$ are continuous at $\boldsymbol{\pi}$ w.r.t. distance $d_{\infty,1}$.

Lemma E.10 (Continuity of $\ddot{Q}$)**.** *Under Cond. E.9, for any $w \in [W]$, any s^w_h, a^w_h and $\boldsymbol{\pi}$, $\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h, a^w_h, \boldsymbol{\pi})$ is continuous at $\boldsymbol{\pi}$ w.r.t. the distance $d_{\infty,1}$ in Def. E.8.*

Proof. The proof is obvious by noting that $\ddot{Q}^{w,\boldsymbol{\pi}}_{\ddot{\boldsymbol{M}},h}(s^w_h, a^w_h, \boldsymbol{\pi})$ is a function resulting from finite multiplication and addition among $\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}(\cdot|\cdot,\cdot,\boldsymbol{\pi})$, $\ddot{r}^w_{\ddot{\boldsymbol{M}},h}(\cdot,\cdot,\boldsymbol{\pi})$ and $\boldsymbol{\pi}$. □

In the following proposition, we will establish the existence of NE based on the existence of strict NE.

Proposition E.11. *Under Cond. E.9, the MT-PAM has at least one NE policy $\boldsymbol{\pi}^{NE}$ satisfying Def. E.5.*

Proof. We first show the mapping $\Gamma^{\text{SNE}}_{\ddot{\boldsymbol{M}}} : \boldsymbol{\Pi} \to \boldsymbol{\Pi}$ is continuous under Cond. E.9. Based on a similar discussion as Lem. E.6 in (Huang et al., 2023),

$$u \to \operatorname*{arg\,max}_{u' \in \Delta(\mathcal{A}^w)} q^\top u' - \|u - u'\|_2^2,$$

is continuous for any fixed $q \in \mathbb{R}^{|\mathcal{A}^w|}$, and

$$q \to \operatorname*{arg\,max}_{u' \in \Delta(\mathcal{A}^w)} q^\top u' - \|u - u'\|_2^2$$

is also continuous for any fixed $u \in \Delta(\mathcal{A}^w)$.

By Lem. E.10, and the rule of composition of continuous functions, $\Gamma^{\text{SNE}}_{\ddot{\boldsymbol{M}}}$ is a continuous mapping. Therefore, $\Gamma^{\text{SNE}}_{\ddot{\boldsymbol{M}}}$ maps from the closed and convex polytope $\boldsymbol{\Pi}$ to a subset of itself. By Brouwers fixed point theorem it has a fixed point. By Lem. E.7, such fixed point is a strict NE of $\ddot{\boldsymbol{M}}$.

Comparing with Def. E.5 and Def. E.6, we know NE is a super-set of strict NE, which implies the existence of NE in the MT-PAM. □

E.2.2. Existence of Nash Equilibrium in MT-MFG as Corollary

Conversion from Multi-Type MF-MDP to MT-PAM Given a Multi-Type MF-MDP $\boldsymbol{M}$, we can convert it to a MT-PAM sharing the same $\{\mu^w_1, \mathcal{S}^w, \mathcal{A}^w, H\}_{w\in[W]}$ with $\boldsymbol{M}$, while the transition and reward functions of $\ddot{\boldsymbol{M}}$ are defined by:

$$\forall w \in [W], \forall h \in [H], \quad \ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}(\cdot|\cdot,\cdot,\boldsymbol{\pi}) := \mathbb{P}^w_{\boldsymbol{M},h}(\cdot|\cdot,\cdot,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h}),\ \ddot{r}_{\ddot{\boldsymbol{M}},h}(\cdot,\cdot,\boldsymbol{\pi}) := r_h(\cdot,\cdot,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h}), \tag{7}$$

where $\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h}$ is the density of agents in all types induced by policy $\boldsymbol{\pi}$ in model $\boldsymbol{M}$ starting from $\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},1} = \boldsymbol{\mu}_1$.

Proposition E.12. *[Existence of NE in MT-MFG] Under Assump. D, the Multi-Type MF-MDP has at least one NE policy $\boldsymbol{\pi}^{NE}$ satisfying Eq. (11).*

Proof. By Lem. E.13, we know the MT-PAM converted from such Multi-Type MF-MDP satisfying Cond. E.9, and by Prop. E.11, the MT-PAM has at least one NE. Easy to check that such NE is also a NE for the Multi-Type MF-MDP satisfying Eq. (11). □

E.2.3. Proofs Related to the MT-PAM converted from Multi-Type MF-MDP

Lemma E.13. *[Lipschitz Continuity of MT-PAM] Given a Multi-Type MF-MDP $\boldsymbol{M}$ satisfying the Lipschitz continuity condition in Assump. D, consider the MT-PAM $\ddot{\boldsymbol{M}}$ converted from $\boldsymbol{M}$ according to Eq. (7), we have $\ddot{\boldsymbol{M}}$ is also Lipschitz continuous that, $\forall w \in [W], h \in [H]$ and any $s^w_h \in \mathcal{S}^w, a^w_h \in \mathcal{A}^w$,*

$$\|\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}(\cdot|s^w_h, a^w_h, \boldsymbol{\pi}) - \ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}(\cdot|s^w_h, a^w_h, \boldsymbol{\pi}')\|_1 \le d_{\infty,1}(\boldsymbol{\pi}, \boldsymbol{\pi}') W \boldsymbol{L}_T \sum_{h'=1}^{h} (1 + \boldsymbol{L}_T)^{h-h'}$$

$$|\ddot{r}^w_{\ddot{\boldsymbol{M}},h}(s^w_h, a^w_h, \boldsymbol{\pi}) - \ddot{r}^w_{\ddot{\boldsymbol{M}},h}(s^w_h, a^w_h, \boldsymbol{\pi}')| \le d_{\infty,1}(\boldsymbol{\pi}, \boldsymbol{\pi}') W \boldsymbol{L}_r \sum_{h'=1}^{h} (1 + \boldsymbol{L}_T)^{h-h'}.$$

Proof. Based on Lem. I.3, as a special case, when $\boldsymbol{M} = \boldsymbol{M}'$, we have:

$$\begin{aligned}\|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M},h}\|_1 \leq &(1+\boldsymbol{L}_T)\|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M},h-1}\|_1 + W \cdot d_{\infty,1}(\boldsymbol{\pi},\boldsymbol{\pi}') \\ =&d_{\infty,1}(\boldsymbol{\pi},\boldsymbol{\pi}')W\sum_{h'=1}^{h}(1+\boldsymbol{L}_T)^{h-h'}.\end{aligned}$$

Therefore, for any $w \in [W]$,

$$\begin{aligned}\|\ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}(\cdot|s^w_h,a^w_h,\boldsymbol{\pi}) - \ddot{\mathbb{P}}^w_{\ddot{\boldsymbol{M}},h}(\cdot|s^w_h,a^w_h,\boldsymbol{\pi}')\|_1 =&\|\mathbb{P}^w_{\boldsymbol{M},h}(\cdot|s^w_h,a^w_h,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h}) - \mathbb{P}^w_{\boldsymbol{M},h}(\cdot|s^w_h,a^w_h,\boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M},h})\|_1 \\ \leq&\boldsymbol{L}_T\|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M},h}\|_1 \leq d_{\infty,1}(\pi,\pi')W\boldsymbol{L}_T\sum_{h'=1}^{h}(1+\boldsymbol{L}_T)^{h-h'}.\end{aligned}$$

and

$$\begin{aligned}|\ddot{r}^w_{\ddot{\boldsymbol{M}},h}(s^w_h,a^w_h,\boldsymbol{\pi}) - \ddot{r}^w_{\ddot{\boldsymbol{M}},h}(s^w_h,a^w_h,\boldsymbol{\pi}')| =&|r^w_h(s^w_h,a^w_h,\mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h}) - r^w_h(s^w_h,a^w_h,\mu^{w,\boldsymbol{\pi}'}_{\boldsymbol{M},h})| \\ \leq&\boldsymbol{L}_r\|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M},h}\|_1 \leq d_{\infty,1}(\boldsymbol{\pi},\boldsymbol{\pi}')W\boldsymbol{L}_r\sum_{h'=1}^{h}(1+\boldsymbol{L}_T)^{h-h'}.\end{aligned}$$

□

F. Missing Details and Proofs for Results in Sec. 4

F.1. Proofs for Lemma and Theorems used for Insights

Lemma 4.3. *[Implication of Local Alignment] Given any $M, \widetilde{M}$ with transition $\mathbb{P}_M$ and $\mathbb{P}_{\widetilde{M}}$, denote $\widehat{\pi}_M^{NE}$ to be an ε_1-approximate NE of M, suppose $d(M, \widetilde{M}|\widehat{\pi}_M^{NE}) \leq \varepsilon_2$, then $\widehat{\pi}_M^{NE}$ is also an $O(\varepsilon_1 + \varepsilon_2)$-approximate NE of $\widetilde{M}$.*

Proof. For any policy $\widetilde{\pi}$, we have:

$$
\begin{aligned}
& J_{\widetilde{M}}(\widetilde{\pi}, \widehat{\pi}_M^{\text{NE}}) - J_{\widetilde{M}}(\widehat{\pi}_M^{\text{NE}}, \widehat{\pi}_M^{\text{NE}}) \\
\leq & J_{\widetilde{M}}(\widetilde{\pi}, \widehat{\pi}_M^{\text{NE}}) - J_M(\widetilde{\pi}, \widehat{\pi}_M^{\text{NE}}) - \left(J_{\widetilde{M}}(\widehat{\pi}_M^{\text{NE}}, \widehat{\pi}_M^{\text{NE}}) - J_M(\widehat{\pi}_M^{\text{NE}}, \widehat{\pi}_M^{\text{NE}})\right) + J_M(\widetilde{\pi}, \widehat{\pi}_M^{\text{NE}}) - J_M(\widehat{\pi}_M^{\text{NE}}, \widehat{\pi}_M^{\text{NE}}) \\
\leq & \varepsilon_1 + J_{\widetilde{M}}(\widetilde{\pi}, \widehat{\pi}_M^{\text{NE}}) - J_M(\widetilde{\pi}, \widehat{\pi}_M^{\text{NE}}) - \left(J_{\widetilde{M}}(\widehat{\pi}_M^{\text{NE}}, \widehat{\pi}_M^{\text{NE}}) - J_M(\widehat{\pi}_M^{\text{NE}}, \widehat{\pi}_M^{\text{NE}})\right) \qquad (\widehat{\pi}_M^{\text{NE}} \text{ is an } \varepsilon_1\text{-NE of } M) \\
\leq & \varepsilon_1 + 2d(M, \widetilde{M}|\widehat{\pi}_M^{\text{NE}}) \\
\leq & \varepsilon_1 + 2\varepsilon_2.
\end{aligned}
$$

□

Theorem 4.4. *[Bridge Policy] If the* `Else`*-branch in Line 6 in Alg. 1 is activated, running Alg. 3 returns a bridge policy $\pi_{Br}^{NE,k}$, such that, $\pi_{Br}^{NE,k}$ is an approximate NE of $M_{Ctr}^{\pi_{Br}^{NE,k}}$.*

Proof. **Thm. 4.4 is just a helper theorem to make it easy for the reader to understand our proofs. It will not be used in the proof of our main results Thm. F.7, so here we only show an informal proof.**

Combining with Lem. E.4 and Thm. E.2, we show the bridge model $\ddot{M}_{\text{Br}}$ has at least one NE $\pi_{\text{Br}}^{\text{NE}}$. In Thm. F.5, we provide upper bound for the distance between the central model of $\pi_{\text{Br}}^{\text{NE}}$ with $\ddot{M}_{\text{Br}}$, which implies $\pi_{\text{Br}}^{\text{NE}}$ is an approximate NE of its central model. □

F.2. Definition of ε-cover of Policy Space

Proposition F.1 (ε-cover of Π)**.** *Consider the set*

$$\Pi_\varepsilon := \{\pi := \{\pi_1, ..., \pi_H\} | \forall h \in [H], s_h \in \mathcal{S}_h,\ \pi_h(\cdot|s_h) \in \mathcal{N}_\varepsilon\},$$

where

$$\mathcal{N}_\varepsilon := \{(\frac{N_1}{N}, ..., \frac{N_A}{N}) | N = \lceil \frac{2A}{\varepsilon} \rceil; N_1, ..., N_A \in \mathbb{N}; \sum_{i=1}^{A} N_i = N\}.$$

Then, Π_ε is an ε-cover of the policy space Π w.r.t. $d_{\infty,1}$ distance.

Proof. For any $u \in \Delta(\mathcal{A})$, there exists a $v \in \mathcal{N}_\varepsilon$, such that, $\|u - v\|_1 \leq \frac{1}{N} \cdot (A-1) + \frac{A-1}{N} \leq \varepsilon$, which implies $\mathcal{N}_\varepsilon$ is an ε-cover of simplex $\Delta(\mathcal{A})$. By definition of Π_ε, we finish the proof. □

F.3. Proofs for Algorithm 2

Theorem F.2 (Adapted from Thm. 4.2 in (Huang et al., 2023))**.** *For any $\delta \in (0, 1)$, during the running of Alg. 2, suppose $M^* \in \bar{\mathcal{M}}$, then w.p. $1 - \delta$, $\forall t \in [T]$, we have $M^* \in \bar{\mathcal{M}}^t$. Besides, denote $\mathbb{H}$ as the hellinger distance, for each $M \in \bar{\mathcal{M}}^t$ with transition $\mathbb{P}_M$ and any $h \in [H]$:*

$$
\begin{aligned}
\sum_{i=1}^{t-1} \mathbb{E}_{\widetilde{\pi}^i, M^*(\pi)}[\mathbb{H}^2(\mathbb{P}_{M,h}(\cdot|s_h^i, a_h^i, \mu_{M,h}^\pi),\ \mathbb{P}_{M^*,h}(\cdot|s_h^i, a_h^i, \mu_{M^*,h}^\pi))] &\leq 2\log(\frac{2|\mathcal{M}|TH}{\delta}) \\
(t-1) \cdot \mathbb{E}_{\pi, M^*(\pi)}[\mathbb{H}^2(\mathbb{P}_{M,h}(\cdot|s_h^i, a_h^i, \mu_{M,h}^\pi),\ \mathbb{P}_{M^*,h}(\cdot|s_h^i, a_h^i, \mu_{M^*,h}^\pi))] &\leq 2\log(\frac{2|\mathcal{M}|TH}{\delta}).
\end{aligned}
$$

Theorem F.3. *Given any reference policy π, $\widetilde{\varepsilon}, \delta \in (0,1)$, $M^* \in \bar{\mathcal{M}}$, if $T = \widetilde{O}(\frac{H^4}{\widetilde{\varepsilon}^2}(\dim_{\mathrm{PE}}(\mathcal{M}, \varepsilon') \wedge (1+L_T)^{2H}(1+L_T H)^2 \dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M}, \varepsilon')) \log^2 \frac{2|\mathcal{M}|TH}{\delta})$ with $\varepsilon' = O(\frac{\widetilde{\varepsilon}}{H^2(1+L_T)^H})$, w.p. $1-\delta$, Alg. 2 terminates at some $T_0 \le T$, and return $\bar{\mathcal{M}}^{T_0}$ s.t. (i) $M^* \in \bar{\mathcal{M}}^{T_0}$ (ii) $\forall M \in \bar{\mathcal{M}}^{T_0}$, $d(M^*, M|\pi) \le \widetilde{\varepsilon}$.*

Proof. Suppose Alg. 2 proceeds to iteration $T_0 \le T$, and does not terminate at Line 6. On the good events in Thm. F.2, we have $M^* \in \bar{\mathcal{M}}^t$ for all $t \le T_0$.

In our first step, we discuss how to provide upper bounds for accumulative model difference depending on two types of P-MBED.

Step 1-(a): Upper Bound Model Difference with Type II P-MBED For any $t \le T_0$, given the fact that $\|P-Q\|_1 \le \sqrt{2}\mathbb{H}(P,Q)$, for any fixed $h \in [H]$, we have:

$$
\begin{aligned}
&\sum_{i=1}^{t-1} \mathbb{E}_{\widetilde{\pi}^i, M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h})\|_1^2] \\
\le& 2\sum_{i=1}^{t-1} \mathbb{E}_{\widetilde{\pi}^i, M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^\pi_{M^t,h})\|_1^2] + 2\sum_{i=1}^{t-1} L_T^2 \|\mu^\pi_{M^*,h} - \mu^\pi_{M^t,h}\|_1^2 \\
\le& 8(1+L_T^2 H^2)\log\frac{2|\mathcal{M}|TH}{\delta}.
\end{aligned} \tag{8}
$$

where in the last step is because, as a result of Lem. I.5, Cauchy's inequality, and $\mathbb{E}^2[X] \le \mathbb{E}[X^2]$, we have:

$$
\begin{aligned}
\sum_{i=1}^{t-1} \|\mu^\pi_{M^*,h} - \mu^\pi_{M^t,h}\|_1^2 \le& (t-1)\cdot H \cdot \mathbb{E}_{\pi,M^*}[\sum_{h'=1}^{h} \|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^\pi_{M^t,h})\|_1^2] \\
\le& 4H^2 \log\frac{2|\mathcal{M}|TH}{\delta}.
\end{aligned}
$$

By Lem. I.1, w.p. $1-\delta/2TH$, for any $t \in [T_0]$ and any $h \in [H]$, we have:

$$
\begin{aligned}
&\sum_{i=1}^{t-1} \|\mathbb{P}_{M^*,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^\pi_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^\pi_{M^*,h})\|_1^2 \\
\le& 96(1+L_T^2H^2)\log\frac{2|\mathcal{M}|TH}{\delta} + C\cdot\log\frac{2TH}{\delta} \le c_1(1+L_T^2H^2)\log\frac{2|\mathcal{M}|TH}{\delta}.
\end{aligned}
$$

for some constant C and c_1. By Lem. D.7, we further have:

$$
\sum_{t=1}^{T_0} \|\mathbb{P}_{M^*,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^\pi_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^\pi_{M^*,h})\|_1 \le c_2((1+L_TH)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0\log\frac{2|\mathcal{M}|TH}{\delta}} + T_0\varepsilon').
$$

for some constant c_2. By Lem. I.1 again, w.p. $1-\delta/2TH$, for any $T_0 \in [T]$ and any $h \in [H]$,

$$
\begin{aligned}
&\sum_{t=1}^{T_0} \mathbb{E}_{\widetilde{\pi}^t, M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h})\|_1] \\
\le& 3c_2((1+L_TH)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0\log\frac{2|\mathcal{M}|TH}{\delta}} + T_0\varepsilon') + C\cdot\log\frac{2TH}{\delta} \\
\le& c_3\cdot((1+L_TH)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon').
\end{aligned}
$$

for some constant c_3. Similarly, we can guarantee by analyzing data collected by (π,π):

$$
\sum_{t=1}^{T_0} \mathbb{E}_{\pi, M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h})\|_1]
$$

$$\leq c_3 \cdot ((1+L_T H)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon').$$

Therefore, by Lem. I.5 again,

$$\begin{aligned}
&\sum_{t=1}^{T_0}\mathbb{E}_{\widetilde{\pi}^t,M^*(\pi)}[\sum_{h=1}^{H}\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^t,h})\|_1] \\
\leq &\sum_{t=1}^{T_0}\mathbb{E}_{\widetilde{\pi}^t,M^*(\pi)}[\sum_{h=1}^{H}\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h})\|_1] + \sum_{t=1}^{T_0}\sum_{h=1}^{H}L_T\|\mu^{\pi}_{M^*,h} - \mu^{\pi}_{M^t,h}\|_1 \\
\leq &c_3\cdot H(1+L_T)^H((1+L_T H)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon').
\end{aligned}$$

where the last step we use:

$$\begin{aligned}
\sum_{t=1}^{T_0}\sum_{h=1}^{H}\|\mu^{\pi}_{M^*,h} - \mu^{\pi}_{M^t,h}\|_1 \leq &\sum_{h=1}^{H}\sum_{h'=1}^{h}(1+L_T)^{h-h'}\sum_{t=1}^{T_0}\mathbb{E}_{\pi,M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h})\|_1] \\
\leq &c_4\cdot H\cdot\frac{(1+L_T)^h-1}{L_T}((1+L_T H)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon').
\end{aligned}$$

Similarly, for model M'^t, we also have:

$$\begin{aligned}
&\sum_{t=1}^{T_0}\sum_{h=1}^{H}\mathbb{E}_{\widetilde{\pi}^t,M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^t,h})\|_1] \\
\leq &c_4\cdot H(1+L_T)^H((1+L_T H)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon').
\end{aligned}$$

Step 1-(b): Upper Bound Model Difference with Type I P-MBED By Thm. F.2 and Lem. I.1, w.p. $1-\delta/2TH$, for any $T_0\in[T]$ and any $h\in[H]$, we have:

$$\sum_{i=1}^{t-1}\|\mathbb{P}_{M^*,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi}_{M^t,h})\|_1^2 \leq c_5\log\frac{2|\mathcal{M}|TH}{\delta} + C\cdot\log\frac{2TH}{\delta} \leq c_6\log\frac{2|\mathcal{M}|TH}{\delta}.$$

As a result of Lem. D.7, we have:

$$\sum_{t=1}^{T_0}\sum_{h=1}^{H}\|\mathbb{P}_{M^*,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^{\pi}_{M^t,h})\|_1 \leq c_7\cdot H(\sqrt{\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0\log\frac{2|\mathcal{M}|TH}{\delta}} + T_0\varepsilon').$$

By Lem. I.1, we have:

$$\begin{aligned}
&\sum_{t=1}^{T_0}\sum_{h=1}^{H}\mathbb{E}_{\widetilde{\pi}^t,M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^{\pi}_{M^t,h})\|_1] \\
&\qquad\leq c_8\cdot H(\sqrt{\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon').
\end{aligned}$$

Similarly, for model M'^t, we also have:

$$\begin{aligned}
&\sum_{t=1}^{T_0}\sum_{h=1}^{H}\mathbb{E}_{\widetilde{\pi}^t,M^*(\pi)}[\|\mathbb{P}_{M^*,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M'^t,h}(\cdot|\widetilde{s}^t_h,\widetilde{a}^t_h,\mu^{\pi}_{M^t,h})\|_1] \\
&\qquad\leq c_8\cdot H(\sqrt{\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon').
\end{aligned}$$

Step 2: Lower Bound on Model Difference On the other hand, since the algorithm does not terminate at step T_0, we have:

$$\begin{aligned}
T_0\widetilde{\varepsilon} &< \sum_{t=1}^{T_0} \mathbb{E}_{\widetilde{\pi}^t, M^t(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M^t,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^t,h}) - \mathbb{P}_{M'^t,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M'^t,h})\|_1] \\
&\leq \sum_{i=1}^{T_0} \mathbb{E}_{\widetilde{\pi}^t, M^*(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M'^t,h}(\cdot|s_h,a_h,\mu^{\pi}_{M'^t,h})\|_1] \\
&\quad + (H+1)\cdot \mathbb{E}_{\widetilde{\pi}^t, M^*(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M^t,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^t,h})\|_1]. \qquad (9)
\end{aligned}$$

where in the last step we apply Lem. I.8. On the one hand, for Type II P-MBED, we have:

$$T_0\widetilde{\varepsilon} \leq c_4 \cdot H(H+2)(1+L_T)^H((1+L_TH)\sqrt{\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon'),$$

by choosing $\varepsilon' \leq \widetilde{\varepsilon}(2c_4H(H+2)(1+L_T)^H)^{-1}$, it implies, for some constant c_8,

$$T_0 \leq c_8 \cdot \frac{H^4(1+L_T)^{2H}(1+L_TH)^2\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')}{\widetilde{\varepsilon}^2}\log^2\frac{2|\mathcal{M}|TH}{\delta}.$$

On the other hand, for Type I P-MBED, we have:

$$T_0\widetilde{\varepsilon} \leq c_7 \cdot H(H+2)(\sqrt{\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')T_0}\log\frac{2|\mathcal{M}|TH}{\delta} + T_0\varepsilon'),$$

by choosing $\varepsilon' \geq \widetilde{\varepsilon}\cdot(c_7H(H+2))^{-1}$, we have:

$$T_0 \leq c_9 \cdot \frac{H^4\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')}{\widetilde{\varepsilon}^2}\log^2\frac{2|\mathcal{M}|TH}{\delta}.$$

As a summary, by choosing

$$T = O(\frac{H^4}{\widetilde{\varepsilon}^2}\min\{\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon'),\ (1+L_T)^{2H}(1+L_TH)^2\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')\}\log^2\frac{2|\mathcal{M}|TH}{\delta})$$

with $\varepsilon' = O(\widetilde{\varepsilon}H^{-2}(1+L_T)^{-H})$, we can guarantee the algorithm will terminates for some $T_0 \leq T$ and return us a model class $\bar{\mathcal{M}}^{T_0}$ satisfying $\max_{M,M'\in\bar{\mathcal{M}}^{T_0}} d(M,M'|\pi) \leq \widetilde{\varepsilon}$, which implies

$$d(M^*, M|\pi) \leq \widetilde{\varepsilon}, \quad \forall M \in \bar{\mathcal{M}}^{T_0}.$$

□

Remark F.4 (Why $(1+L_T)^H$ Disappears if Considering Type-I P-MBED?)**.** *From the proof above, especially the proof in Step 1-(a) and Step 1-(b), we can see that during the model elimination, what matters is the model distance conditioning on the density induced by the corresponding models, i.e.* $\|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h}) - \mathbb{P}_{M^*,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^*,h})\|_1$*. Therefore, if we consider the Type-*I *P-MBED, we do not need additional conversion between* $\|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h}) - \mathbb{P}_{M^*,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^*,h})\|_1$ *and* $\|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^*,h}) - \mathbb{P}_{M',h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M^*,h})\|_1$*, which is the origin of the exponential term* $(1+L_T)^H$ *in the upper bound regarding Type-*II *P-MBED.*

F.4. Proofs for Algorithm 3

Recall the notations for central models in Appx. E.1.

Theorem F.5. *Suppose we feed Alg. 3 with a model class* $\ddot{\mathcal{M}}$*, the bridge model* $\ddot{M}_{Br}$ *it computes is a valid model, and by choosing* $\bar{\varepsilon} = \varepsilon_0/\min\{2HL_r\frac{(1+L_T)^H-1}{L_T}, 2H(H+1)((1+L_T)^H-1)\}$*, for any reference policy* π *and its associated central model* $\ddot{M}^{\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}})$*, we have:*

$$\max_{\widetilde{\pi}} \mathbb{E}_{\widetilde{\pi},\ddot{M}^{\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}})(\pi)}[\sum_{h=1}^{H}\|\ddot{\mathbb{P}}_{\ddot{M}^{\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}}),h}(\cdot|s_h,a_h,\pi) - \ddot{\mathbb{P}}_{Br,h}(\cdot|s_h,a_h,\pi)\|_1] \leq (H+3)\varepsilon_0,$$

$$\max_{\widetilde{\pi}} \mathbb{E}_{\widetilde{\pi}, \ddot{M}^{\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}})(\pi)}[\sum_{h=1}^{H} |\ddot{r}_{\ddot{M}^{\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}}),h}(s_h, a_h, \pi) - \ddot{r}_{Br,h}(s_h, a_h, \pi)|] \leq L_r H(H+4)\varepsilon_0.$$

Proof. In the proof, for notation simplicity, given the refernce policy π, we use $\ddot{M}^{\pi}_{\text{Ctr}}$ as a short note of $\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\pi;\ddot{\mathcal{M}})$, i.e. the central model regarding π.

Validity of Construction First of all, note that $\Pi_{\bar{\varepsilon}}$ is an $\bar{\varepsilon}$-cover of the policy space. Therefore, for any π, there must exist at least one $\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}$ satisfying $d_{\infty,1}(\pi, \widetilde{\pi}) \leq \bar{\varepsilon}$ which ensures $\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ > 0$. So the transition and reward functions in the bridge model is well-defined, and also continuous in π w.r.t. distance $d_{\infty,1}$.

Upper Bound on Transition Difference By definition,

$$\begin{aligned}
\|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{\text{Br},h}(\cdot|s_h, a_h, \pi)\|_1 =& \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi) - \frac{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})}{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+}\|_1 \\
\leq& \frac{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1}{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+}
\end{aligned}$$

We only need to care about those $\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}$ with $[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ > 0$, i.e. $d_{\infty,1}(\pi, \widetilde{\pi}) < 2\bar{\varepsilon}$. Given the condition when Alg. 1 call Alg. 3, we have $\mathcal{B}^{\varepsilon_0}_{\pi}(\ddot{M}^{\pi}_{\text{Ctr}};\ddot{\mathcal{M}}) > \frac{|\ddot{\mathcal{M}}|}{2}$ for any π. Therefore, for any π and $\widetilde{\pi}$ with $d_{\infty,1}(\pi, \widetilde{\pi}) \leq 2\bar{\varepsilon}$, there exists a model $\ddot{M}_{\text{share}}$ such that $\ddot{M}_{\text{share}} \in \mathcal{B}^{\varepsilon_0}_{\pi}(\ddot{M}^{\pi}_{\text{Ctr}};\ddot{\mathcal{M}}) \cap \mathcal{B}^{\varepsilon_0}_{\widetilde{\pi}}(\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}};\ddot{\mathcal{M}})$, which implies for any π'

$$\begin{aligned}
&\mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1] \\
\leq& \mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \pi)\| + \sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1 \\
&+ \sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi}) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1] \\
\leq& \varepsilon_0 + 2H((1+L_T)^H - 1)\bar{\varepsilon} + \mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi}) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1]. \qquad \text{(Lem. E.4)}
\end{aligned}$$

where by applying Lem. I.7, we have:

$$\begin{aligned}
&\mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi}) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1] \\
\leq& \mathbb{E}_{\pi', \ddot{M}_{\text{share}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi}) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1] \\
&+ H \cdot \mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \pi)\|_1] \\
\leq& \mathbb{E}_{\pi', \ddot{M}_{\text{share}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi}) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1] + H\varepsilon_0 \qquad (\ddot{M}_{\text{share}} \in \mathcal{B}^{\varepsilon_0}_{\pi}(\ddot{M}^{\pi}_{\text{Ctr}};\ddot{\mathcal{M}})) \\
\leq& \mathbb{E}_{\pi', \ddot{M}_{\text{share}}(\widetilde{\pi})}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi}) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1] \\
&+ H \cdot \mathbb{E}_{\pi', \ddot{M}_{\text{share}}(\widetilde{\pi})}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{\ddot{M}_{\text{share}},h}(\cdot|s_h, a_h, \widetilde{\pi})\|_1] + H\varepsilon_0
\end{aligned}$$

$$\leq (H+1)\varepsilon_0 + 2H^2((1+L_T)^H - 1)\bar{\varepsilon}. \qquad (\ddot{M}_{\text{share}} \in \mathcal{B}^{\varepsilon_0}_{\widetilde{\pi}}(\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}}; \ddot{\mathcal{M}}); \text{Lem. E.4})$$

which implies,

$$\begin{aligned}
&\mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h,a_h,\pi) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h,a_h,\widetilde{\pi})\|_1] \\
\leq& \varepsilon_0 + 2H((1+L_T)^H - 1)\bar{\varepsilon} + (H+1)\varepsilon_0 + 2H^2((1+L_T)^H - 1)\bar{\varepsilon} \\
\leq& (H+2)\varepsilon_0 + 2H(H+1)((1+L_T)^H - 1)\bar{\varepsilon} \\
\leq& (H+3)\varepsilon_0. \qquad (2H(H+1)((1+L_T)^H - 1)\bar{\varepsilon} \leq \varepsilon_0)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\forall \pi, \pi', \quad \mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h,a_h,\pi) - \ddot{\mathbb{P}}_{\text{Br},h}(\cdot|s_h,a_h,\pi)\|_1] \leq& \frac{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+ (H+3)\varepsilon_0}{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+} \\
\leq& (H+3)\varepsilon_0.
\end{aligned}$$

Upper Bound on Reward Difference By definition, for each h, s_h, a_h, we have:

$$\begin{aligned}
&|\ddot{r}_{\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\pi;\ddot{\mathcal{M}}),h}(s_h,a_h,\pi) - \ddot{r}_{\text{Br},h}(s_h,a_h,\pi)| \\
=& |\ddot{r}_{\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\pi;\ddot{\mathcal{M}}),h}(s_h,a_h,\pi) - \frac{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+ \ddot{r}_{\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\widetilde{\pi};\ddot{\mathcal{M}}),h}(s_h,a_h,\widetilde{\pi})}{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+}| \\
=& \frac{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+ |\ddot{r}_{\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\pi;\ddot{\mathcal{M}}),h}(s_h,a_h,\pi) - \ddot{r}_{\ddot{M}^{\varepsilon_0}_{\text{Ctr}}(\widetilde{\pi};\ddot{\mathcal{M}}),h}(s_h,a_h,\widetilde{\pi})|}{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+}.
\end{aligned}$$

Similarly, for those $\widetilde{\pi} \in \Pi_{\bar{\varepsilon}}$ with $[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+ > 0$, we have:

$$\begin{aligned}
&|\ddot{r}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(s_h,a_h,\pi) - \ddot{r}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(s_h,a_h,\widetilde{\pi})| \\
=& |r_h(s_h,a_h,\mu^{\pi}_{M^{\pi}_{\text{Ctr}},h}) - r_h(s_h,a_h,\mu^{\widetilde{\pi}}_{M^{\widetilde{\pi}}_{\text{Ctr}},h})| \\
\leq& L_r \|\mu^{\pi}_{M^{\pi}_{\text{Ctr}},h} - \mu^{\widetilde{\pi}}_{M^{\widetilde{\pi}}_{\text{Ctr}},h}\|_1 \\
\leq& L_r H d_{\infty,1}(\pi,\pi') + L_r \mathbb{E}_{\pi, M^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h'=1}^{h} \|\mathbb{P}_{M^{\pi}_{\text{Ctr}},h'}(\cdot|s_{h'-1},a_{h'-1},\mu^{\pi}_{M^{\pi}_{\text{Ctr}},h'-1}) - \mathbb{P}_{M^{\widetilde{\pi}}_{\text{Ctr}},h'}(\cdot|s_{h'-1},a_{h'-1},\mu^{\widetilde{\pi}}_{M^{\widetilde{\pi}}_{\text{Ctr}},h'-1})\|_1] \\
\leq& 2L_r H\bar{\varepsilon} + L_r \mathbb{E}_{\pi, \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(\cdot|s_h,a_h,\pi) - \ddot{\mathbb{P}}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(\cdot|s_h,a_h,\widetilde{\pi})\|_1] \\
\leq& L_r(H+4)\varepsilon_0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\forall \pi, \pi', \quad \mathbb{E}_{\pi', \ddot{M}^{\pi}_{\text{Ctr}}(\pi)}[\sum_{h=1}^{H} |\ddot{r}_{\ddot{M}^{\pi}_{\text{Ctr}},h}(s_h,a_h,\pi) - \ddot{r}_{\ddot{M}^{\widetilde{\pi}}_{\text{Ctr}},h}(s_h,a_h,\widetilde{\pi})|] \leq& \frac{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+ H(H+4)\varepsilon_0}{\sum_{\widetilde{\pi}\in\Pi_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi,\widetilde{\pi})]^+} \\
\leq& L_r H(H+4)\varepsilon_0.
\end{aligned}$$

□

Next we prove an important Lemma based on results in theorem above, which indicates that the bridge policy constructed in Alg. 3 is close to the NE of its central model.

Lemma F.6. *Suppose the* `Else`*-branch in Line 6 if activated in Alg. 2, for policy $\pi^{NE,k}_{Br}$ and its corresponding central model $M^k_{Ctr} := \arg\max_{M\in\mathcal{M}^k} |\mathcal{B}^{\varepsilon_0}_{\pi^{NE,k}_{Br}}(M;\mathcal{M}^k)|$, we have:*

$$\mathcal{E}^{NE}_{M^k_{Ctr}}(\pi^{NE,k}_{Br}) := \max_{\pi} \Delta_{M^k_{Ctr}}(\pi, \pi^{NE,k}_{Br}) \leq 2(1+L_r)(H+4)\varepsilon_0.$$

Proof. For any policy π, we have

$$
\begin{aligned}
&\Delta_{M^k_{\text{Ctr}}}(\pi, \pi^{\text{NE},k}_{\text{Br}}) \\
\leq &\Delta_{\ddot{M}^k_{\text{Ctr}}}(\pi, \pi^{\text{NE},k}_{\text{Br}}) - \Delta_{\ddot{M}_{\text{Br}}}(\pi, \pi^{\text{NE},k}_{\text{Br}}) && (\Delta_{\ddot{M}_{\text{Br}}}(\pi, \pi^{\text{NE},k}_{\text{Br}}) \leq 0) \\
\leq &|\ddot{J}_{\ddot{M}^k_{\text{Ctr}}}(\pi, \pi^{\text{NE},k}_{\text{Br}}) - \ddot{J}_{\ddot{M}_{\text{Br}}}(\pi, \pi^{\text{NE},k}_{\text{Br}})| + |\ddot{J}_{\ddot{M}^k_{\text{Ctr}}}(\pi^{\text{NE},k}_{\text{Br}}, \pi^{\text{NE},k}_{\text{Br}}) - \ddot{J}_{\ddot{M}_{\text{Br}}}(\pi^{\text{NE},k}_{\text{Br}}, \pi^{\text{NE},k}_{\text{Br}})| \\
\leq &\mathbb{E}_{\pi^{\text{NE},k}_{\text{Br}}, \ddot{M}^k_{\text{Ctr}}(\pi^{\text{NE},k}_{\text{Br}})}[\sum_{h=1}^{H} |\ddot{r}_{\ddot{M}^k_{\text{Ctr}},h}(s_h, a_h, \pi^{\text{NE},k}_{\text{Br}}) - \ddot{r}_{\ddot{M}_{\text{Br}},h}(s_h, a_h, \pi^{\text{NE},k}_{\text{Br}})| \\
&\qquad + \|\ddot{\mathbb{P}}_{\ddot{M}^k_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi^{\text{NE},k}_{\text{Br}}), \ddot{\mathbb{P}}_{\ddot{M}_{\text{Br}},h}(\cdot|s_h, a_h, \pi^{\text{NE},k}_{\text{Br}})\|_1] \\
&+ \mathbb{E}_{\pi, \ddot{M}^k_{\text{Ctr}}(\pi^{\text{NE},k}_{\text{Br}})}[\sum_{h=1}^{H} |\ddot{r}_{\ddot{M}^k_{\text{Ctr}},h}(s_h, a_h, \pi^{\text{NE},k}_{\text{Br}}) - \ddot{r}_{\ddot{M}_{\text{Br}},h}(s_h, a_h, \pi^{\text{NE},k}_{\text{Br}})| \\
&\qquad + \|\ddot{\mathbb{P}}_{\ddot{M}^k_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi^{\text{NE},k}_{\text{Br}}), \ddot{\mathbb{P}}_{\ddot{M}_{\text{Br}},h}(\cdot|s_h, a_h, \pi^{\text{NE},k}_{\text{Br}})\|_1] \\
\leq &2 \max_{\pi} \mathbb{E}_{\pi, \ddot{M}^k_{\text{Ctr}}(\pi^{\text{NE},k}_{\text{Br}})}[\sum_{h=1}^{H} |\ddot{r}_{\ddot{M}^k_{\text{Ctr}},h}(s_h, a_h, \pi^{\text{NE},k}_{\text{Br}}) - \ddot{r}_{\ddot{M}_{\text{Br}},h}(s_h, a_h, \pi^{\text{NE},k}_{\text{Br}})|] \\
&+ 2 \max_{\pi} \mathbb{E}_{\pi, \ddot{M}^k_{\text{Ctr}}(\pi^{\text{NE},k}_{\text{Br}})}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}^k_{\text{Ctr}},h}(\cdot|s_h, a_h, \pi^{\text{NE},k}_{\text{Br}}), \ddot{\mathbb{P}}_{\ddot{M}_{\text{Br}},h}(\cdot|s_h, a_h, \pi^{\text{NE},k}_{\text{Br}})\|_1] \\
\leq &2(1 + L_r H)(H + 4)\varepsilon_0. && \text{(Thm. F.5)}
\end{aligned}
$$

which finishes the proof. □

F.5. Proofs for Algorithm 1

Theorem 4.5. *In Alg. 1, by choosing $\varepsilon_0 = \frac{\varepsilon}{8(1+L_r H)(H+4)}$, $\widetilde{\varepsilon} = \frac{\varepsilon_0}{6}$, and choosing $\bar{\varepsilon}$ according to Thm. F.5, w.p. $1-\delta$, (1) if the* `If-Branch` *in Line 5 is activated: we have $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$; (2) otherwise, in the* `Else-Branch` *in Line 6: either we return the $\pi^{NE,k}_{Br}$ which is an ε-approximate NE for M^*; or the algorithm continues with $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$.*

Proof. We separately discuss the if and else branches in the algorithm.

Proof for `If-Branch` in Line 5 On the events in Thm. F.3, for any $\widetilde{M} \notin \mathcal{B}^{\varepsilon_0}_{\pi^k}(M^*; \mathcal{M}^k)$, we have $d(M^*, \widetilde{M}|\pi^k) \geq \varepsilon_0 > \widetilde{\varepsilon}$, which implies $\widetilde{M} \notin \mathcal{M}^{k+1}$. Combining the condition of `If-Branch`, we have:

$$
|\mathcal{M}^{k+1}| \leq |\mathcal{B}^{\varepsilon_0}_{\pi^k}(M^*; \mathcal{M}^k)| \leq \frac{|\mathcal{M}^k|}{2}.
$$

Proof for `Else-Branch` in Line 6 First of all, on the events in Thm. F.3, we have $d(M^*, \widetilde{M}^k|\pi^{\text{NE},k}_{\text{Br}}) \leq \widetilde{\varepsilon}$. By applying Lem. I.2, it implies:

$$
\begin{aligned}
&|\Delta_{M^*}(\pi, \pi^{\text{NE},k}_{\text{Br}}) - \Delta_{\widetilde{M}^k}(\pi, \pi^{\text{NE},k}_{\text{Br}})| \\
\leq &\mathbb{E}_{\pi, M^*(\pi^{\text{NE},k}_{\text{Br}})}[\sum_{h=1}^{H} \|\mathbb{P}_{M^*,h}(\cdot|s_h, a_h, \mu^{\pi^{\text{NE},k}_{\text{Br}}}_{M^*,h}) - \mathbb{P}_{\widetilde{M}^k,h}(\cdot|s_h, a_h, \mu^{\pi^{\text{NE},k}_{\text{Br}}}_{\widetilde{M}^k,h})\|_1] \\
&+ (2L_r H + 1)\mathbb{E}_{\pi^{\text{NE},k}_{\text{Br}}, M^*(\pi^{\text{NE},k}_{\text{Br}})}[\sum_{h=1}^{H} \|\mathbb{P}_{M^*,h}(\cdot|s_h, a_h, \mu^{\pi^{\text{NE},k}_{\text{Br}}}_{M^*,h}) - \mathbb{P}_{\widetilde{M}^k,h}(\cdot|s_h, a_h, \mu^{\pi^{\text{NE},k}_{\text{Br}}}_{\widetilde{M}^k,h})\|_1] \\
\leq &2(L_r H + 1)\widetilde{\varepsilon}.
\end{aligned}
$$

Also note that:

$$
\Delta_{M^*}(\pi, \pi^{\text{NE},k}_{\text{Br}}) = \Delta_{M^*}(\pi, \pi^{\text{NE},k}_{\text{Br}}) - \Delta_{\widetilde{M}^k}(\pi, \pi^{\text{NE},k}_{\text{Br}}) + \Delta_{\widetilde{M}^k}(\pi, \pi^{\text{NE},k}_{\text{Br}}).
$$

In the following, we separately discuss two cases.

Case 1: $\mathcal{E}^{\mathbf{NE}}_{\widetilde{M}^k}(\pi^{\mathbf{NE},k}_{\mathbf{Br}}) \leq \frac{3\varepsilon}{4}$ and Line 9 is activated Given that $\widetilde{\varepsilon} \leq \frac{\varepsilon}{16(1+L_rH)}$:

$$\begin{aligned}\Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) \leq& |\Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) - \Delta_{\widetilde{M}^k}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}})| + \mathcal{E}^{\mathrm{NE}}_{\widetilde{M}^k}(\pi^{\mathrm{NE},k}_{\mathrm{Br}}) \qquad (\mathcal{E}^{\mathrm{NE}}_{\widetilde{M}^k}(\pi^{\mathrm{NE},k}_{\mathrm{Br}}) = \max_\pi \Delta_{\widetilde{M}^k}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}))\\ \leq& 2(L_rH+1)\widetilde{\varepsilon} + \frac{3\varepsilon}{4} \leq \varepsilon.\end{aligned}$$

which implies $\pi^{\mathrm{NE},k}_{\mathrm{Br}}$ is an ε-NE of M^*.

Case 2: $\mathcal{E}^{\mathbf{NE}}_{\widetilde{M}^k}(\pi^{\mathbf{NE},k}_{\mathbf{Br}}) > \frac{3\varepsilon}{4}$ and Line 9 is not activated As a result, for any policy π,

$$\Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) \geq -|\Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) - \Delta_{\widetilde{M}^k}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}})| + \Delta_{\widetilde{M}^k}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) \geq \Delta_{\widetilde{M}^k}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) - 2(L_rH+1)\widetilde{\varepsilon}.$$

Therefore, by our choice of $\widetilde{\varepsilon}$,

$$\max_\pi \Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) \geq \mathcal{E}^{\mathrm{NE}}_{\widetilde{M}^k}(\pi^{\mathrm{NE},k}_{\mathrm{Br}}) - 2(L_rH+1)\widetilde{\varepsilon} \geq \frac{5\varepsilon}{8}.$$

On the other hand, by Lem. F.6, for any π, we have:

$$\begin{aligned}&\Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) - 2(1+L_rH)(H+4)\varepsilon_0\\ \leq& |\Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}})| - |\Delta_{M^k_{\mathrm{Ctr}}}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}})| \qquad \text{(Here we apply Lem. F.6)}\\ \leq& |\Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) - \Delta_{M^k_{\mathrm{Ctr}}}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}})|\\ \leq& \mathbb{E}_{\pi, M^*(\pi^{\mathrm{NE},k}_{\mathrm{Br}})}[\sum_{h=1}^H \|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi^{\mathrm{NE},k}_{\mathrm{Br}}}_{M^*,h}) - \mathbb{P}_{M^k_{\mathrm{Ctr}},h}(\cdot|s_h,a_h,\mu^{\pi^{\mathrm{NE},k}_{\mathrm{Br}}}_{M^k_{\mathrm{Ctr}},h})\|_1]\\ &+ (2L_rH+1)\mathbb{E}_{\pi^{\mathrm{NE},k}_{\mathrm{Br}}, M^*(\pi^{\mathrm{NE},k}_{\mathrm{Br}})}[\sum_{h=1}^H \|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi^{\mathrm{NE},k}_{\mathrm{Br}}}_{M^*,h}) - \mathbb{P}_{M^k_{\mathrm{Ctr}},h}(\cdot|s_h,a_h,\mu^{\pi^{\mathrm{NE},k}_{\mathrm{Br}}}_{M^k_{\mathrm{Ctr}},h})\|_1]\\ \leq& (2L_rH+2)d(M^*, M^k_{\mathrm{Ctr}}|\pi^{\mathrm{NE},k}_{\mathrm{Br}}).\end{aligned}$$

According to the choice of ε_0, we have $2(1+L_rH)(H+4)\varepsilon_0 \leq \frac{\varepsilon}{4}$, therefore,

$$d(M^*, M^k_{\mathrm{Ctr}}|\pi^{\mathrm{NE},k}_{\mathrm{Br}}) \geq \frac{1}{2L_rH+2}\Big(\max_\pi \Delta_{M^*}(\pi, \pi^{\mathrm{NE},k}_{\mathrm{Br}}) - 2(1+L_rH)(H+4)\varepsilon_0\Big) \geq \frac{3\varepsilon}{16(L_rH+1)}.$$

Next we try to show that models in $\mathcal{B}^{\varepsilon_0}_{\pi^{\mathrm{NE},k}_{\mathrm{Br}}}(M^k_{\mathrm{Ctr}}, \mathcal{M}^k)$ will be eliminated. For any $M \in \mathcal{B}^{\varepsilon_0}_{\pi^{\mathrm{NE},k}_{\mathrm{Br}}}(M^k_{\mathrm{Ctr}}, \mathcal{M}^k)$, because of $\widetilde{\varepsilon} < \frac{\varepsilon}{48(L_rH+1)}$ we have:

$$d(M, M^*|\pi^{\mathrm{NE},k}_{\mathrm{Br}}) \geq d(M^k_{\mathrm{Ctr}}, M^*|\pi^{\mathrm{NE},k}_{\mathrm{Br}}) - d(M, M^k_{\mathrm{Ctr}}|\pi^{\mathrm{NE},k}_{\mathrm{Br}}) \geq \frac{3\varepsilon}{16(L_rH+1)} - \varepsilon_0 \geq \frac{\varepsilon}{16(L_rH+1)} > \widetilde{\varepsilon}.$$

On the event in Thm. F.3 (which holds with probability $1-\delta$), $M \notin \mathcal{M}^{k+1}$, which implies,

$$|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k| - |\mathcal{B}^{\varepsilon_0}_{\pi^{\mathrm{NE},k}_{\mathrm{Br}}}(M^k_{\mathrm{Ctr}}, \mathcal{M}^k)| \leq |\mathcal{M}^k|/2.$$

□

Theorem F.7. *[Sample Complexity of Learning MFGs] Under Assump. A and B, by running Alg. 1 with Alg. 2 as* `ModelElim` *and Alg. 3 as* `BridgePolicy`*, and hyper-parameter choices according to Thm. F.3, 4.5, and F.5, w.p. $1-\delta$, Alg. 1 will terminate at some $k \leq \log_2|\mathcal{M}|+1$ and return us an ε-NE of M^*, and the number of trajectories consumed is at most $O(\frac{H^7}{\varepsilon^2}(1+L_r)^2(\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon') \wedge H^3(1+L_TH)^2(1+L_T)^{2H}\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon'))\log^3\frac{|\mathcal{M}|}{\delta})$ where $\varepsilon' = O(\varepsilon/H^3(1+L_r)(1+L_T)^H)$, and in $\widetilde{O}$ we omit logarithmic terms of $\varepsilon, H, \log|\mathcal{M}|, \dim_{\mathrm{PE}}, 1+L_T$ and $1+L_r$.*

Proof. As a result of Thm. 4.5, w.p. $1 - \frac{\delta}{\log_2|\mathcal{M}|+1} \cdot (\log_2|\mathcal{M}|+1) = 1-\delta$, there exists a step $k \le \log_2|\mathcal{M}|+1$ such that Alg. 1 will terminate the return us an ε-approximate NE of M^*. The total number of trajectories required is:

$$
\begin{aligned}
&(\log_2|\mathcal{M}|+1)\cdot 2HT\\
=&(\log_2|\mathcal{M}|+1)\cdot \widetilde{O}(\frac{H^5}{\widetilde{\varepsilon}^2}(\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')\wedge(1+L_T)^{2H}(1+L_TH)^2\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon'))\log^2\frac{2|\mathcal{M}|TH}{\delta})\\
=&\widetilde{O}(\frac{H^7}{\varepsilon^2}(1+L_rH)^2\Big(\dim_{\mathrm{PE}}(\mathcal{M},\varepsilon')\wedge H^3(1+L_TH)^2(1+L_T)^{2H}\dim^{\mathrm{II}}_{\mathrm{PE}}(\mathcal{M},\varepsilon')\Big)\log^3|\mathcal{M}|).
\end{aligned}
$$

where we use the fact that by Thm. F.3, we choose $\widetilde{\varepsilon} = \frac{\varepsilon_0}{6} = O(\frac{\varepsilon}{(1+L_rH)H})$, and $\varepsilon' = O(\widetilde{\varepsilon}/H^2(1+L_T)^H) = O(\varepsilon/H^3(1+L_rH)(1+L_T)^H)$. □

F.6. Sample Complexity Separation between Mean-Filed Control and Mean-Field Games

In this section, we establish the separation between of RL in MFC and MFGs from information theoretical perspective.

A Basic Recap of the MFC Setting In MFC, similar to single-agent RL, we are interested in finding a policy $\widehat{\pi}^*_{\mathrm{Opt}}$ to approximately minimize the optimality gap $\mathcal{E}_{\mathrm{Opt}}(\pi) := \max_{\widetilde{\pi}} J_{M^*}(\widetilde{\pi};\boldsymbol{\mu}^{\widetilde{\pi}}_{M^*}) - J_{M^*}(\pi;\boldsymbol{\mu}^{\pi}_{M^*})$, i.e.,

$$\mathcal{E}_{\mathrm{Opt}}(\widehat{\pi}^*_{\mathrm{Opt}}) \le \varepsilon. \tag{10}$$

Exponential Lower Bound in Tabular RL for Mean-Field Control Our results are based on a different query model from Def. 2.1 defined below.

Definition F.8 (Strong Query Model)**.** The Strong Query Model (SQM) can take a policy π and return a sequence of transition function $\{\mathbb{P}^{\pi}_h(\cdot|\cdot,\cdot)\}_{h=1}^H$, such that $\mathbb{P}^{\pi}_h(\cdot|s_h,a_h) := \mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu^{\pi}_{M^*,h})$ for any $h\in[H]$, $s_h\in\mathcal{S}_h$, $a_h\in\mathcal{A}_h$.

The SQM is strictly stronger than the sample query model in Def. 2.1, because given the conditional model $\{\mathbb{P}^{\pi}_h(\cdot|\cdot,\cdot)\}_{h=1}^H$, one can sample arbitrary trajectories by arbitrary policies from it, and therefore, recover the data collection process in Def. 2.1. In the following, we investigate the number of SQM queries required to identify ε-optimal policy in MFC setting. We show that, under Assump. A and B, even in the tabular setting, MFC requires queries exponential to the number of states and actions.

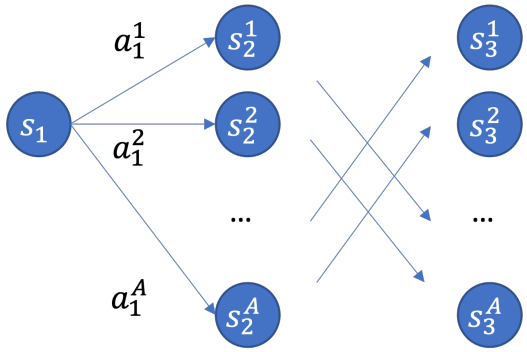

Figure 1. Construction of Lower Bound

Theorem F.9. *[Exponential Lower Bound for MFC] Given arbitrary $L_T > 0$ and $d \ge 2$, consider tabular MF-MDPs satisfying Assump. B with Lipschitz coefficient L_T, $|\mathcal{S}| = |\mathcal{A}| = d$ and $H = 3$. For any algorithm Alg, and any $\varepsilon \le \frac{L_T}{d+1}$, there exists an MDP M^* and a model class $\mathcal{M}$ satisfying $M^*\in\mathcal{M}$, and $|\mathcal{M}| = \Omega((\frac{L_T}{d\varepsilon})^{d-1})$, s.t., if Alg only queries GM or DCP for at most K times with $K \le |\mathcal{M}|/2-1$, the probability that Alg produces an ε-optimal policy is less than $1/2$.*

Proof. Our proof is divided into three parts: construction of hard MF-MDP instance, construction of model class $\mathcal{M}$, and the proof of lower bound.

Part 1: Construction of Hard Examples We construct a three layer MDP as shown in Fig. 1. The initial state distribution is fixed to be $\mu_1(s_1) = 1$, and we have S states and A actions available at each layer with $S = A = d$. The transition at initial state is deterministic, i.e., $\mathbb{P}(s_2^i|s_1, a_1^i, \mu_1) = 1$. At the second layer, given $L_T \leq 1$, there exists an optimal state density μ_2^*, such that, $\forall i \in [S], j \in [A]$ and $\forall \mu_2 \in \Delta(\mathcal{S})$:

$$\mathbb{P}(s_3^1|s_2^i, a_2^j, \mu_2) = \frac{1}{2} + 2\varepsilon \cdot \Big[1 - \frac{L_T}{4\varepsilon}\|\mu_2 - \mu_2^*\|_1\Big]^+, \quad \mathbb{P}(s_3^2|s_2^i, a_2^j, \mu_2) = \frac{1}{2} - 2\varepsilon \cdot \Big[1 - \frac{L_T}{4\varepsilon}\|\mu_2 - \mu_2^*\|_1\Big]^+.$$

where $[x]^+ = \max\{x, 0\}$. As for the reward function, we have zero reward at each state action in the previous two layers, and for the third layer, we have only have non-zero reward at $r_3(s_3^1, \cdot, \cdot) = 1$ and $r_3(s_3^i, \cdot, \cdot) = 0$ for all $i \neq 1$.

As we can see, for arbitrary policy π, we have $\mu_2^\pi(s_2^i) = \pi(a_1^i|s_1)$. Besides, the optimal policy should be taking action to make sure $\mu_2 = \mu_2^*$, which can be achieved by setting $\pi^*(a_1^i|s_1) = \mu_2^*(s_2^i)$, and then take arbitrary policy at the second layer. Even if the agent just wants to achieve ε-near-optimal policy, it at least has to determine the position of set $\{\mu : \|\mu - \mu_2^*\|_1 \leq \frac{4\varepsilon}{L_T}\}$. The key difficulty here is to explore and gather information which can be used to infer μ_2^*.

We further reduce the difficulty of the exploration by providing for the learner with the transition at initial state and the third layer (or equivalently, the available representation function for the first and third layers is unique) and all the information of reward function. All the learner need to do is to identify the correct feature for the second layer and use it to obtain the optimal policy (at the initial state) to maximize the return.

Next, we verify the above model belongs to the low-rank Mean-Field MDP. For $h = 1$, it's easy to see $\mathbb{P}(s_2^i|s_1, a_1^j, \mu_1) = \phi_1(s_1, a_1^j, \mu_1)^\top \psi_1(s_2^i)$, where $\phi_1(s_1, a_1^j, \mu_1) = \mathbf{e}_j$ and $\psi_1(s_2^i) = \mathbf{e}_i$, and $\mathbf{e}_{(\cdot)}$ is the one-hot vector with the $(\cdot)$-th element equal 1. For the second layer, given a density $\mu \in \Delta(\mathcal{S})$, we use ϕ_{μ,L_T} to denote the following feature function class that, $\forall i \in [S], j \in [A], \mu' \in \Delta(\mathcal{S})$,

$$\phi_{\mu,L_T}(s_2^i, a_2^j, \mu') := (\frac{1}{2} + 2\varepsilon \cdot \Big[1 - \frac{L_T}{4\varepsilon}\|\mu' - \mu\|_1\Big]^+, \frac{1}{2} - 2\varepsilon \cdot \Big[1 - \frac{L_T}{4\varepsilon}\|\mu' - \mu\|_1\Big]^+, 0, .., 0)^\top \in \mathbb{R}^d.$$

and the next state feature function is $\psi(s_3^i) = \mathbf{e}_i^\top, \quad \forall i \in [d]$. It's easy to verify that the transition can be decomposed to $\phi_{\mu_2^*,L_T}(\cdot, \cdot, \mu_2)^\top \psi(s_3^i)$, and the above feature satisfies the normalization property:

$$\|\sum_{i\in[d]} \psi(s_3^i) g(s_3^i)\| \leq \sqrt{2}d, \quad \forall g : \mathcal{S} \to \{-1, 1\}.$$

Besides, we verify that for any choice of μ, the induced transition function is L_T-Lipschitz:

$$\begin{aligned}
&\|\mathbb{P}_{\mu,L_T}(\cdot|s_2^i, a_2^j, \mu') - \mathbb{P}_{\mu,L_T}(\cdot|s_2^i, a_2^j, \mu'')\|_1 \\
=&\sum_{l\in[S]} |\phi_{\mu,L_T}(s_2^i, a_2^j, \mu')^\top \psi(s_3^l) - \phi_{\mu,L_T}(s_2^i, a_2^j, \mu'')\psi(s_3^l)| \\
=&2 \cdot 2\varepsilon|\Big[1 - \frac{L_T}{4\varepsilon}\|\mu - \mu''\|_1\Big]^+ - \Big[1 - \frac{L_T}{4\varepsilon}\|\mu - \mu'\|_1\Big]^+| \\
\leq&L_T|\|\mu - \mu'\|_1 - \|\mu - \mu''\|_1| \leq L_T\|\mu' - \mu''\|_1
\end{aligned}$$

Part 2: Construction of Model Class Given an integer ζ, we denote $\mathcal{N}_\zeta := \{\mu|\mu(s_2^i) = N(s_2^i)/\zeta,\ N(s_2^i) \in \mathbf{N},\ \sum_{i\in[S]} N(s_2^i) = \zeta\}$. In another word, $\mathcal{N}_\zeta$ includes all state density with resolution $1/\zeta$. Now, consider $\mathcal{N}_{\lfloor \frac{L_T}{5\varepsilon} \rfloor}$. For each $\mu, \mu' \in \mathcal{N}_{\lfloor \frac{L_T}{5\varepsilon} \rfloor}$, we should have:

$$\|\mu - \mu'\|_1 \geq 2/\lfloor \frac{L_T}{5\varepsilon} \rfloor \geq \frac{10\varepsilon}{L_T} > \frac{8\varepsilon}{L_T}.$$

Therefore, if we consider the set $\mathcal{B}(\mu, \frac{4\varepsilon}{L_T}) := \{\mu' \in \Delta(\mathcal{S})|\|\mu - \mu'\|_1 \leq \frac{4\varepsilon}{L_T}\}$, we can expect $\mathcal{B}(\mu, \frac{4\varepsilon}{L_T}) \cap \mathcal{B}(\mu', \frac{4\varepsilon}{L_T}) = \emptyset$ for any $\mu, \mu' \in \mathcal{N}_{\lfloor \frac{L_T}{5\varepsilon} \rfloor}$. Given arbitrary $N \leq |\mathcal{N}_{\lfloor \frac{L_T}{5\varepsilon} \rfloor}| = \frac{(\lfloor \frac{L_T}{5\varepsilon} \rfloor + d-1)!}{(\lfloor \frac{L_T}{5\varepsilon} \rfloor)!(d-1)!} = \Omega((\frac{L_T}{d\varepsilon})^{d-1})$, we can find $N - 1$ different elments $\{\mu_2^1, ..., \mu_2^N\} \subset \mathcal{N}_{\lfloor \frac{L_T}{5\varepsilon} \rfloor}$ and construct (here we only specify the representation at the second layer, since we assume the other layers are known)

$$\mathcal{M}^{[N]} := \{M^n := (\phi_{\mu_2^n,L_T}, \psi)|n \in [N]\}.$$

For analysis, we introduce another model $\bar{M}$ which shares the transition and reward function as M^ns but for the transition of second layer, it has:

$$\mathbb{P}(s_3^1|s_2^i, a_2^j, \mu_2) = \mathbb{P}(s_3^2|s_2^i, a_2^j, \mu_2) = \frac{1}{2}, \quad \forall i \in [S], j \in [A], \mu_2 \in \Delta(\mathcal{S}).$$

We define:

$$\bar{\phi}(\cdot,\cdot,\cdot) = (\frac{1}{2}, ..., \frac{1}{2}) \in \mathbb{R}^d.$$

and define:

$$\mathcal{M} := \mathcal{M}^{[N]} \cup \{(\bar{\phi}, \psi)\}.$$

Note that $\bar{M} = (\bar{\phi}, \psi) \in \mathcal{M}$.

Part 3: Establishing Lower Bound Now, we consider the following learning setting: the environment randomly select one model M from $\mathcal{M}$ and provide the entire representation feature class $\mathcal{M}$ (which is also the entire model class) to the learner; then, the learner can repeatedly use gathered information to compute a policy π^k and query it with SQM for each iteration, and output a final policy after K steps. We want to show that, for arbitrary algorithm, there exists at least one model in $\mathcal{M}$ which cost number of queries linear w.r.t. N before identifying the optimal policy.

In the following, we use $\mathcal{E}_{k,M^n}$ to denote the event that in the first k trajectories, there is at least one policy (or equivalently, density μ_2^π) used to query SQM resulting in $\|\mu_2^\pi - \mu^n\|_1 \leq \frac{4\varepsilon}{L_T}$. The key observation is that, given arbitrary algorithm Alg, for arbitrary fixed $n \in [N]$, if Alg never deploy a policy π (or equivalently, query an density μ_2^π) satisfying $\|\mu_2^\pi - \mu^n\|_1 \leq \frac{4\varepsilon}{L_T}$, the algorithm can not distinguish between M^n and $\bar{M}$, and should behave similar in both M^n and $\bar{M}$. Therefore,

$$\Pr_{M^n,\text{Alg}}(\mathcal{E}_{k,M^n}^{\complement}) = \Pr_{\bar{M},\text{Alg}}(\mathcal{E}_{k,M^n}^{\complement}), \quad \forall k \in [K].$$

which also implies:

$$\Pr_{M^n,\text{Alg}}(\mathcal{E}_{k,M^n}) = \Pr_{\bar{M},\text{Alg}}(\mathcal{E}_{k,M^n}), \quad \forall k \in [K].$$

We use $\text{Alg}(K)$ to denote the policy output by the algorithm in the final. Besides, we use $\Pi(\mu, b_0) := \{\pi | \|\mu_2^\pi - \mu\|_1 \leq b_0\}$ to denote the set of policies, which can lead to a density μ_2^π close to μ. Then, we have:

$$\begin{aligned}
&\sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})) - \Pr_{\bar{M},\text{Alg}}(\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})) \\
=& \sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\{\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})\} \cap \{\mathcal{E}_{K,M^n}\}) - \Pr_{\bar{M},\text{Alg}}(\{\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})\} \cap \{\mathcal{E}_{K,M^n}\}) \\
&+ \sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\{\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})\} \cap \{\mathcal{E}_{K,M^n}^{\complement}\}) - \Pr_{\bar{M},\text{Alg}}(\{\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})\} \cap \{\mathcal{E}_{K,M^n}^{\complement}\}) \\
=& \sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\{\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})\} \cap \{\mathcal{E}_{K,M^n}\}) - \Pr_{\bar{M},\text{Alg}}(\{\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})\} \cap \{\mathcal{E}_{K,M^n}\}) \\
\leq& \sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\mathcal{E}_{k,M^n}) \Big(\Pr_{M^n,\text{Alg}}(\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T}) | \mathcal{E}_{K,M^n}) - \Pr_{\bar{M},\text{Alg}}(\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T}) | \mathcal{E}_{K,M^n}) \Big) \\
\leq& \sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\mathcal{E}_{k,M^n}) = \sum_{n\in[N]} \Pr_{\bar{M},\text{Alg}}(\mathcal{E}_{k,M^n}) \leq K.
\end{aligned}$$

where the last step is because,

$$\sum_{n\in[N]} \Pr_{\bar{M},\text{Alg}}(\mathcal{E}_{k,M^n}) \leq \sum_{n\in[N]} \sum_{k=1}^{K} \Pr_{\bar{M},\text{Alg}}(\|\mu_2^{\pi^k} - \mu^n\|_1 \leq \frac{4\varepsilon}{L_T}) = \sum_{k=1}^{K} \sum_{n\in[N]} \Pr_{\bar{M},\text{Alg}}(\|\mu_2^{\pi^k} - \mu^n\|_1 \leq \frac{4\varepsilon}{L_T}) \leq \sum_{k=1}^{K} 1 = K.$$

$$(\mathcal{B}(\mu^i, \tfrac{4\varepsilon}{L_T}) \cap \mathcal{B}(\mu^j, \tfrac{4\varepsilon}{L_T}) = \emptyset \text{ for all } i \neq j)$$

Therefore, the average success probability would be:

$$\begin{aligned}
&\Pr(M=\bar{M}) + \sum_{n\in[N]} \Pr(\{M=M^n\}\cap\{\text{Alg}(K)\in\Pi(\mu^n,\frac{4\varepsilon}{L_T})\}) && \text{(Each policy is optimal in } \bar{M}.\text{)}\\
=&\frac{1}{|\mathcal{M}|} + \frac{1}{|\mathcal{M}|}\sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\text{Alg}(K)\in\Pi(\mu^n,\frac{4\varepsilon}{L_T})) \le \frac{K+1}{|\mathcal{M}|}.
\end{aligned}$$

As a result, even if $K = \frac{|\mathcal{M}|}{2} - 1 = O(N)$, there exists $n \in [N]$, such that, the failure rate

$$\Pr_{M^n,\text{Alg}}(\text{Alg}(K) \notin \mathcal{B}(\pi^*_{M^n}, \frac{4\varepsilon}{L_T})) \ge \frac{1}{2}.$$

□

G. Proofs for Multi-Type MFGs

G.1. More Details about the Setting

In Multi-Type MF-MDP, we will denote $\mu_h^w \in \Delta(\mathcal{S}^w)$ to be a state density at step h in type w, and define $\mu^w := \{\mu_h^w\}_{h\in[H]}$ to be the collection over all h. For the policies, we define $\Pi^w := \{\pi^w := \{\pi_h^w\} | \forall h \in [H], \pi_h^w : \mathcal{S}_h^w \to \Delta(\mathcal{A}_h^w)\}$, and $\mathbf{\Pi} := \{\boldsymbol{\pi} := \{\pi^w\}_{w\in[W]} | \forall w \in [W], \pi^w \in \Pi^w\}$. In this paper, we only consider policies in $\mathbf{\Pi}$, i.e. the set of non-stationary Markovian policies.

In order to distinguish with (single-type) MF-MDP setting, for notations regarding the collection of densities or policies over all groups, we use the bold font, i.e. $\boldsymbol{\mu}_h := \{\mu_h^w\}_{w\in[W]}$ and $\boldsymbol{\mu} := \{\boldsymbol{\mu}_h\}_{h\in[W]}$, $\boldsymbol{\pi} := \{\pi^w\}_{w\in[W]}$ and $\boldsymbol{\pi}_h := \{\pi_h^w\}_{w\in[W]}$. When a policy $\boldsymbol{\pi}$ and a model $\boldsymbol{M}$ is speicified, we use $\boldsymbol{\mu}_{\boldsymbol{M}}^{\boldsymbol{\pi}} := \{\mu_{\boldsymbol{M}}^{w,\boldsymbol{\pi}}\}_{w\in[W]} = \{\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}}\}_{h\in[H]}$ to denote the collection of densities of W groups induced by the policy π in model $\boldsymbol{M}$, where $\mu_{\boldsymbol{M}}^{w,\boldsymbol{\pi}} := \{\mu_{\boldsymbol{M},h}^{w,\boldsymbol{\pi}}\}_{h\in[H]}$ and $\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}} := \{\mu_{\boldsymbol{M},h}^{w,\boldsymbol{\pi}}\}_{w\in[W]}$. When a policy $\boldsymbol{\pi} \in \mathbf{\Pi}$ is specified, the evolution of the densities in all groups can be described by:

$$\forall h \in [H],\ \forall w \in [W], \quad \mu_{\boldsymbol{M},h+1}^{w,\boldsymbol{\pi}} = \Gamma_{\boldsymbol{M},h}^{w,\pi^w}(\mu_{\boldsymbol{M},h}^{w,\boldsymbol{\pi}}),$$
$$\text{with } \Gamma_{\boldsymbol{M},h}^{w,\pi^w}(\mu_h^w)(\cdot) := \sum_{s_h^w, a_h^w} \mu_h^w(s_h^w)\pi_h^w(a_h^w|s_h^w)\mathbb{P}_{\boldsymbol{M},h}^w(\cdot|s_h^w, a_h^w, \boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}})$$

Similarly to MF-MDP setting, given two policies $\widetilde{\boldsymbol{\pi}}, \boldsymbol{\pi} \in \mathbf{\Pi}$, we can define the value functions for each group following $\widetilde{\boldsymbol{\pi}}$ while conditioning on $\boldsymbol{\pi}$:

$$Q_{\boldsymbol{M},h}^{w,\widetilde{\boldsymbol{\pi}}}(\cdot,\cdot;\boldsymbol{\mu}_{\boldsymbol{M}}^{\boldsymbol{\pi}}) := \mathbb{E}_{\widetilde{\boldsymbol{\pi}},\boldsymbol{M}(\boldsymbol{\pi})}[\sum_{h'=h}^{H} r_{h'}^w(s_{h'}^w, a_{h'}^w, \boldsymbol{\mu}_{\boldsymbol{M},h'}^{\boldsymbol{\pi}})|s_h^w = \cdot, a_h^w = \cdot],$$
$$V_{\boldsymbol{M},h}^{w,\widetilde{\boldsymbol{\pi}}}(\cdot;\boldsymbol{\mu}_{\boldsymbol{M}}^{\boldsymbol{\pi}}) := \mathbb{E}_{\widetilde{\boldsymbol{\pi}},\boldsymbol{M}(\boldsymbol{\pi})}[\sum_{h'=h}^{H} r_{h'}^w(s_{h'}^w, a_{h'}^w, \boldsymbol{\mu}_{\boldsymbol{M},h'}^{\boldsymbol{\pi}})|s_h^w = \cdot],$$
$$J_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}};\boldsymbol{\pi}) := \mathbb{E}_{s_1^w\sim\mu_1^w}[V_{\boldsymbol{M},1}^{w,\widetilde{\boldsymbol{\pi}}}(s_1^w)].$$

where we use $\mathbb{E}_{\widetilde{\boldsymbol{\pi}},\boldsymbol{M}(\boldsymbol{\pi})}$ to denote the expectation over trajectories generated by executing policy $\widetilde{\boldsymbol{\pi}}$ in $\boldsymbol{M}$ conditioning on $\boldsymbol{\pi}$, i.e. the transitions $\mathbb{P}_{\boldsymbol{M},h}^w(\cdot|\cdot,\cdot,\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}})$ and rewards $r_h^w(\cdot,\cdot,\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}})$ are fixed by $\boldsymbol{\pi}$. Besides, we denote $J_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}};\boldsymbol{\pi}) := \mathbb{E}_{s_1^w\sim\mu_1^w}[V_{\boldsymbol{M},1}^{w,\boldsymbol{\pi}}(s_1^w;\boldsymbol{\mu}_{\boldsymbol{M}}^{\boldsymbol{\pi}})]$ to be the expected return of type w in model $\boldsymbol{M}$ by executing $\widetilde{\boldsymbol{\pi}}$ conditioning on $\boldsymbol{\pi}$. The Nash Equilibrium policy in Multi-Type MFG is defined to be the policy $\boldsymbol{\pi}^{\text{NE}} := \{\pi^{w,\text{NE}}\}_{w\in[W]}$ satisfying:

$$\forall w \in [W],\ \forall \widetilde{\boldsymbol{\pi}} \in \mathbf{\Pi}, \quad J_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}};\boldsymbol{\pi}^{\text{NE}}) \le J_{\boldsymbol{M}}^w(\boldsymbol{\pi}^{\text{NE}};\boldsymbol{\pi}^{\text{NE}}). \tag{11}$$

We define $\Delta_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}},\boldsymbol{\pi}^w) := J_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}};\boldsymbol{\pi}) - J_{\boldsymbol{M}}^w(\boldsymbol{\pi};\boldsymbol{\pi})$, and define $\mathcal{E}_{\boldsymbol{M}}^{w,\text{NE}}(\boldsymbol{\pi}) := \max_{\widetilde{\boldsymbol{\pi}}} \Delta_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}},\boldsymbol{\pi})$. Our goal in this setting is to find an ε-approximate NE policy $\widehat{\boldsymbol{\pi}}^{\text{NE}} := \{\widehat{\boldsymbol{\pi}}^{w,\text{NE}}\}_{w\in[W]}$ such that:

$$\forall w \in [W], \quad \mathcal{E}_{\boldsymbol{M}^*}^{w,\text{NE}}(\widehat{\boldsymbol{\pi}}) \le \varepsilon. \tag{12}$$

G.2. Conversion from MT-MFG to MFG with Constrained Policy Space

Intuitively, the construction is made by integrating the state and action spaces, which will result in a MFG with transition and reward functions following some block diagnoal structure.

Given a MT-MFG $\boldsymbol{M} := \{(\mu_1^w, H, \mathcal{S}^w, \mathcal{A}^w, \mathbb{P}_{\boldsymbol{M}}^w, r^w)_{w\in[W]}\}$, we denote the converted MF-MDP by $M_{\text{MFG}} := \{\mu_1, H, \mathcal{S}_{\text{MFG}}, \mathcal{A}_{\text{MFG}}, \mathbb{P}_{\text{MFG}}, r_{\text{MFG}}\}$, where we have the extended state space $\mathcal{S}_{\text{MFG}} := \bigcup_{w\in[W]}(\mathcal{S}^w \times \{w\})$ and action space $\mathcal{A}_{\text{MFG}} := \bigcup_{w\in[W]}(\mathcal{A}^w \times \{w\})$. As we can see, the new state/action space is the collection of all states/actions agumented by the group index $w \in [W]$. In this way, states and actions in different groups can be distinguished by the group index w. Next, we construct a new initial distribution $\mu_1 := [\frac{\mu_1^1}{W}, \frac{\mu_1^2}{W}, ..., \frac{\mu_1^W}{W}]$ by concatenating all the initial distributions with normalization. For the policy, we define

$$\Pi^\dagger := \{\pi | \forall w \in [W], \pi(a^w \circ w | s^w \circ w) = \pi^w(a^w \circ w | s^w \circ w),\ \text{for some } \pi^w \in \Pi^w\}, \tag{13}$$

with $\Pi^w := \{\pi^w : \mathcal{S}^w \to \Delta(\mathcal{A}^w)\}$. In another word, $\Pi^\dagger$ includes and only includes policies taking actions sharing the same group index with states, and we only consider the policies $\pi \in \Pi^\dagger$.

G.2.1. Definition of Transition/Reward Functions in the Lifted MF-MDP

Next, given a density $\mu_h := [\frac{\mu_h^1}{W}, ..., \frac{\mu_h^W}{W}] \in \Delta(\mathcal{S}_{\text{MFG}})$ with $\mu_h^w \in \Delta(\mathcal{S}^w)$, the transition and reward functions in the converted MFG is defined by (note that by definition of $\Pi^\dagger$, we only need to consider the case when the state and action share the group index):

$$\mathbb{P}_{\text{MFG},h}(s_{h+1}^{\widetilde{w}} \circ \widetilde{w} | s_h^w \circ w, a_h^w \circ w, \mu_h) = \begin{cases} \mathbb{P}_{\boldsymbol{M},h}^w(s_{h+1}^w | s_h^w, a_h^w, \mu_h^w), & \text{if } \widetilde{w} = w \\ 0, & \text{otherwise.} \end{cases}$$
$$r_{\text{MFG},h}(s_h^w \circ w, a_h^w \circ w, \mu_h) = r_h^w(s_h^w, a_h^w, \mu_h^w).$$

For the sake of rigor, we include the definition for the transition/reward functions on those $s_h^w \circ w$ and $a_h^{\widetilde{w}} \circ \widetilde{w}$ with $w \neq \widetilde{w}$. We define $\mathbb{P}_{\text{MFG},h}(\cdot | s_h^w \circ w, a_h^{\widetilde{w}} \circ \widetilde{w}, \mu_h)$ to be a uniform distribution over $\mathcal{S}_{\text{MFG}}$, and $r_{\text{MFG},h}(s_h^w \circ w, a_h^{\widetilde{w}} \circ \widetilde{w}, \mu_h) = 0$, for any $\mu_h \in \Delta(\mathcal{S}_{\text{MFG}})$.

After specifying a policy $\pi \in \Pi^\dagger$, denote $\boldsymbol{\pi} := \{\pi^w\}_{w \in [W]}$ to be the MT-MFG policy that π corresponds to, we can verify that the state density $\mu_{M_{\text{MFG}},h}^\pi \in \Delta(\mathcal{S}_{\text{MFG}})$ evolves according to:

$$\forall h \in [H]: \quad \mu_{M_{\text{MFG}},h}^\pi \leftarrow [\frac{\mu_{\boldsymbol{M},h}^{1,\boldsymbol{\pi}}}{W}, ..., \frac{\mu_{\boldsymbol{M},h}^{W,\boldsymbol{\pi}}}{W}]. \tag{14}$$

where recall $\mu_{\boldsymbol{M},h}^{w,\boldsymbol{\pi}}$ denotes the density of type w induced by $\boldsymbol{\pi}$ in model $\boldsymbol{M}$. To see this, by induction,

$$\begin{aligned}
\mu_{M_{\text{MFG}},h}^\pi(s^w \circ w) &= \sum_{s_{h-1} \in \mathcal{S}_{\text{MFG}}, a_{h-1} \in \mathcal{A}_{\text{MFG}}} \mu_{M_{\text{MFG}},h}^\pi(s_{h-1}) \pi(a_{h-1} | s_{h-1}) \mathbb{P}_{\text{MFG},h}(s_h^w \circ w | s_{h-1}, a_{h-1}, \mu_{M_{\text{MFG}},h-1}^\pi) \\
&= \sum_{s_{h-1}^w \in \mathcal{S}^w, a_{h-1}^w \in \mathcal{A}^w} \mu_{M_{\text{MFG}},h-1}^{w,\pi}(s_{h-1}^w \circ w) \cdot \pi(a_{h-1}^w \circ w | s_{h-1}^w \circ w) \mathbb{P}_{\boldsymbol{M},h}^w(s_h^w | s_{h-1}^w, a_{h-1}^w, \mu_{M_{\text{MFG}},h-1}^{w,\pi}), \\
&= \sum_{s_{h-1}^w \in \mathcal{S}^w, a_{h-1}^w \in \mathcal{A}^w} \mu_{\boldsymbol{M},h-1}^{w,\boldsymbol{\pi}}(s_{h-1}^w)/W \cdot \pi^w(a_{h-1}^w | s_{h-1}^w) \mathbb{P}_{\boldsymbol{M},h}^w(s_h^w | s_{h-1}^w, a_{h-1}^w, \mu_{M_{\text{MFG}},h-1}^{w,\pi}), \\
&= \mu_{\boldsymbol{M},h}^{w,\boldsymbol{\pi}}(s^w)/W.
\end{aligned}$$

Intuitively, in the converted MFG, following a policy $\pi \in \Pi^\dagger$, if an agent starts from the initial state with index w, it will follow a trajectory as if it is generated in the original MT-MFG. In the following, we will call M_{MFG} (or $\boldsymbol{M}$) the corresponding MFG (or MT-MFG) of $\boldsymbol{M}$ (or M_{MFG}).

G.3. Assumptions and Additional Definitions

Recall the definition of $\{\mathcal{M}^w\}_{w \in [W]}$ and $\boldsymbol{\mathcal{M}}$ discussed in Sec. 2, In the following, we use $\mathcal{M}_{\text{MFG}}$ to denote the model class including MFG models converted from models in $\boldsymbol{\mathcal{M}}$ according to the method discussed in Appx. G.2, and denote M_{MFG}^* to be the one converted from $\boldsymbol{M}^*$.

We have the following assumptions, which can be regarded as a generalization of Assump. A, B and Def. 2.1.

Assumption C (Realizability)**.** The true model $\boldsymbol{M}^* \in \boldsymbol{\mathcal{M}}$.

Assumption D (Lipschitz Continuity for MT-MFG)**.** For any $\boldsymbol{M} \in \boldsymbol{\mathcal{M}}$, and for two arbitrary policies $\boldsymbol{\pi}, \widetilde{\boldsymbol{\pi}}$

$$\begin{aligned}
\forall w \in [W],\ \forall h, s^w, a^w, \quad \|\mathbb{P}_{\boldsymbol{M},h}^w(\cdot | s_h^w, a_h^w, \boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}}) - \mathbb{P}_{\boldsymbol{M},h}^w(\cdot | s_h^w, a_h^w, \boldsymbol{\mu}_{\boldsymbol{M},h}^{\widetilde{\boldsymbol{\pi}}})\|_1 &\leq \frac{\boldsymbol{L}_T}{W} \|\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}} - \boldsymbol{\mu}_{\boldsymbol{M},h}^{\widetilde{\boldsymbol{\pi}}}\|_1, \\
|r_h^w(s_h^w, a_h^w, \boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}}) - r_h^w(s_h^w, a_h^w, \boldsymbol{\mu}_{\boldsymbol{M},h}^{\widetilde{\boldsymbol{\pi}}})| &\leq \frac{\boldsymbol{L}_T}{W} \|\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}} - \boldsymbol{\mu}_{\boldsymbol{M},h}^{\widetilde{\boldsymbol{\pi}}}\|_1.
\end{aligned}$$

Here we introduce a normalization factor W given that $\|\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}} - \boldsymbol{\mu}_{\boldsymbol{M},h}^{\widetilde{\boldsymbol{\pi}}}\|_1 = \sum_{w \in [W]} \|\mu_{\boldsymbol{M},h}^{w,\boldsymbol{\pi}} - \mu_{\boldsymbol{M},h}^{w,\widetilde{\boldsymbol{\pi}}}\|_1$.

Definition G.1 (Trajectory Sampling Model in MT-MFG)**.** The learner can query the sampling model with an arbitrary policy $\boldsymbol{\pi} := \{\pi^1, ..., \pi^W\}$, a group index w and another policy $\widetilde{\boldsymbol{\pi}} := \{\widetilde{\pi}^1, ..., \widetilde{\pi}^W\}$, and receive a trajectory by executing $\widetilde{\pi}^w$ while the transition and reward functions are fixed by $\boldsymbol{\pi}$, i.e. $\mathbb{P}_{\boldsymbol{M}^*,h}^w(\cdot | \cdot, \cdot, \boldsymbol{\mu}_{\boldsymbol{M}^*,h}^{\boldsymbol{\pi}})$ and $r_h^w(\cdot, \cdot, \boldsymbol{\mu}_{\boldsymbol{M}^*,h}^{\boldsymbol{\pi}})$.

Similar to the sampling model in Def. 2.1, the model above can be implemented by utilizing the observation of an individually deviating agent with type w following policy $\widetilde{\pi}^w$ while the other agents follows $\boldsymbol{\pi}$ in a large Multi-Type MARL system.

Moreover, for learning in the lifted MFGs, note that a sampling model in M^*_{MFG} as described in Def. 2.1 can be implemented by Def. G.1. To see this, given two policies $\pi, \widetilde{\pi} \in \Pi^\dagger$, which correspond to $\boldsymbol{\pi} := \{\pi^1, ..., \pi^W\}$ and $\widetilde{\boldsymbol{\pi}} := \{\widetilde{\pi}^1, ..., \widetilde{\pi}^W\}$, respectively, the trajectory can be generated by first uniformly sample $w \in [W]$, and then sample a trajectory with Def. G.1 with $\boldsymbol{\pi}$ converted from π, type w and policy $\widetilde{\pi}^w$.

Proposition G.2. *Given a MT-MFG model class $\boldsymbol{\mathcal{M}}$ satisfying Assump. D, consider its converted MF-MDP model class $\mathcal{M}_{MFG}$, for any $M \in \mathcal{M}_{MFG}$, and any $\pi, \widetilde{\pi} \in \Pi^\dagger$, we have,*

$$\forall h \in [H], \quad \|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^{\pi}_{M,h}) - \mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^{\widetilde{\pi}}_{M,h})\|_1 \leq \boldsymbol{L}_T \cdot \|\mu^{\pi}_{M,h} - \mu^{\widetilde{\pi}}_{M,h}\|_1. \tag{15}$$

$$|r_h(s_h, a_h, \mu^{\pi}_{M,h}) - r_h(s_h, a_h, \mu^{\widetilde{\pi}}_{M,h})| \leq \boldsymbol{L}_r \cdot \|\mu^{\pi}_{M,h} - \mu^{\widetilde{\pi}}_{M,h}\|_1 \tag{16}$$

Proof. According to the definition in Appx. G.2.1, for those s_h, a_h with different group index, their transition or reward differences will be 0. Therefore, we only need to consider the case when s_h, a_h share the group index.

As we explained in Eq. (14), given $\pi, \widetilde{\pi} \in \Pi^\dagger$, which corresponds to $\boldsymbol{\pi}, \widetilde{\boldsymbol{\pi}} \in \boldsymbol{\Pi}$, respectively, we have:

$$\|\mu^{\pi}_{M,h} - \mu^{\widetilde{\pi}}_{M,h}\|_1 = \frac{1}{W} \sum_{w \in [W]} \|\mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h}, \mu^{w,\widetilde{\boldsymbol{\pi}}}_{\boldsymbol{M},h}\|_1,$$

where $\boldsymbol{M}$ is the corresponding MT-MFG model of M. Combining with Assump. D, we finish the proof. □

G.4. Constrained Nash Equilibrium

Proposition 5.1. *Given a MT-MFG $\boldsymbol{M}$ and its lifted MFG M_{MFG}, we have: (1) an ε-constrained NE $\widehat{\pi}^{NE}_{Cstr} \in \Pi^\dagger$ for M_{MFG} is a $(W\varepsilon)$-NE in $\boldsymbol{M}$; (2) an ε-NE $\widehat{\boldsymbol{\pi}}^{NE}$ in $\boldsymbol{M}$ is an ε-constrained NE for M_{MFG}.*

Proof. Given any $\pi \in \Pi^\dagger$, we denote its corresponding policy in MT-MFG by $\boldsymbol{\pi} := \{\pi^1, ..., \pi^W\}$ with $\pi^w : \mathcal{S}^w \to \Delta(\mathcal{A}^w)$ and $\pi(a^w_h \circ w | s^w_h \circ w) = \pi^w(a^w_h | s^w_h)$. Conversely, given any $\boldsymbol{\pi} := \{\pi^1, ..., \pi^W\}$, we can convert it to a policy in $\Pi^\dagger$, which we denote by π. For $\widehat{\pi}^{\text{NE}}_{\text{Cstr}}$, we denote its correspondence in MT-MFG by $\widehat{\boldsymbol{\pi}}^{\text{NE}} := \{\widehat{\pi}^{\text{NE},1}, ...\widehat{\pi}^{\text{NE},W}\}$.

Note that, given any $\pi, \widetilde{\pi} \in \Pi^\dagger$ and their correspondence $\boldsymbol{\pi} := \{\pi^1, ...\pi^W\}$ and $\widetilde{\boldsymbol{\pi}} := \{\widetilde{\pi}^1, ...\widetilde{\pi}^W\}$, we have:

$$J_{M_{\text{MFG}}}(\widetilde{\pi}; \pi) = \mathbb{E}_{\widetilde{\pi}; M_{\text{MFG}}(\pi)}[\sum_{h=1}^{H} r_{M_{\text{MFG}},h}(s_h, a_h, \mu^{\pi}_{M_{\text{MFG}},h})] = \frac{1}{W} \sum_{w=1}^{W} \mathbb{E}_{\widetilde{\pi}^w; \boldsymbol{M}(\boldsymbol{\pi})}[\sum_{h=1}^{H} r^w_h(s^w_h, a^w_h, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h})].$$

where recall r^w is the reward in type w in MT-MFG and $\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h} := \{\mu^{w,\pi}_{\boldsymbol{M},h}\}_{w \in [W]}$ is the collection of densities for all groups. Consider the case when $\boldsymbol{\pi} = \widehat{\boldsymbol{\pi}}^{\text{NE}}$ and $\widetilde{\pi}^{\widetilde{w}} \leftarrow \widehat{\pi}^{\text{NE},\widetilde{w}}$ for all $\widetilde{w}$ except $\widetilde{w} = w$, we have:

$$\begin{aligned}
\varepsilon \geq & J_{M_{\text{MFG}}}(\widetilde{\pi}; \widehat{\pi}^{\text{NE}}_{\text{Cstr}}) - J_{M_{\text{MFG}}}(\widehat{\pi}^{\text{NE}}_{\text{Cstr}}; \widehat{\pi}^{\text{NE}}_{\text{Cstr}}) \\
= & \frac{1}{W}\Big(\mathbb{E}_{\widetilde{\pi}^w; \boldsymbol{M}(\widehat{\boldsymbol{\pi}}^{\text{NE}})}[\sum_{h=1}^{H} r^w_h(s^w_h, a^w_h, \boldsymbol{\mu}^{\widehat{\boldsymbol{\pi}}^{\text{NE}}}_{\boldsymbol{M},h})] - \mathbb{E}_{\widehat{\pi}^{\text{NE},w}; \boldsymbol{M}(\widehat{\boldsymbol{\pi}}^{\text{NE}})}[\sum_{h=1}^{H} r^w_h(s^w_h, a^w_h, \boldsymbol{\mu}^{\widehat{\boldsymbol{\pi}}^{\text{NE}}}_{\boldsymbol{M},h})]\Big) \\
= & \frac{1}{W}\Big(J^w_{\boldsymbol{M}}(\widetilde{\boldsymbol{\pi}}, \widehat{\boldsymbol{\pi}}^{\text{NE}}) - J^w_{\boldsymbol{M}}(\widetilde{\boldsymbol{\pi}}, \widehat{\boldsymbol{\pi}}^{\text{NE}})\Big).
\end{aligned}$$

By repeating such discussion for any $w \in [W]$ and any π^w, we complete the proof for argument (1).

On the other hand, given an ε-approximate NE $\widehat{\boldsymbol{\pi}}^{\text{NE}}$ in $\boldsymbol{M}$ and its corresponding $\widehat{\pi}^{\text{NE}}_{\text{Cstr}}$ in M_{MFG}, for any $\pi \in \Pi^\dagger$ we have:

$$\begin{aligned}
& J_{M_{\text{MFG}}}(\pi; \widehat{\pi}^{\text{NE}}_{\text{Cstr}}) - J_{M_{\text{MFG}}}(\widehat{\pi}^{\text{NE}}_{\text{Cstr}}; \widehat{\pi}^{\text{NE}}_{\text{Cstr}}) \\
= & \frac{1}{W} \sum_{w \in [W]} \Big(\mathbb{E}_{\pi^w; \boldsymbol{M}(\widehat{\boldsymbol{\pi}}^{\text{NE}})}[\sum_{h=1}^{H} r^w_h(s^w_h, a^w_h, \boldsymbol{\mu}^{\widehat{\boldsymbol{\pi}}^{\text{NE}}}_{\boldsymbol{M},h})] - \mathbb{E}_{\widehat{\pi}^{\text{NE},w}; \boldsymbol{M}(\widehat{\boldsymbol{\pi}}^{\text{NE}})}[\sum_{h=1}^{H} r^w_h(s^w_h, a^w_h, \boldsymbol{\mu}^{\widehat{\boldsymbol{\pi}}^{\text{NE}}}_{\boldsymbol{M},h})]\Big).
\end{aligned}$$

To upper bound the RHS, for each $w \in [W]$, we consider an arbitrary policy $\widetilde{\boldsymbol{\pi}}$ with with $\widetilde{\pi}^{\widetilde{w}} = \widehat{\pi}^{\mathrm{NE},\widetilde{w}}$ for all $\widetilde{w}$ except $\widetilde{\pi}^w = \pi^w$, we should have:

$$\mathbb{E}_{\pi^w;\boldsymbol{M}(\widehat{\boldsymbol{\pi}}^{\mathrm{NE}})}[\sum_{h=1}^{H} r_h^w(s_h^w, a_h^w, \boldsymbol{\mu}_{\boldsymbol{M},h}^{\widehat{\boldsymbol{\pi}}^{\mathrm{NE}}})] - \mathbb{E}_{\widehat{\pi}^{\mathrm{NE},w};\boldsymbol{M}(\widehat{\boldsymbol{\pi}}^{\mathrm{NE}})}[\sum_{h=1}^{H} r_h^w(s_h^w, a_h^w, \boldsymbol{\mu}_{\boldsymbol{M},h}^{\widehat{\boldsymbol{\pi}}^{\mathrm{NE}}})] = J_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}}, \widehat{\boldsymbol{\pi}}^{\mathrm{NE}}) - J_{\boldsymbol{M}}^w(\widetilde{\boldsymbol{\pi}}, \widehat{\boldsymbol{\pi}}^{\mathrm{NE}}) \leq \varepsilon,$$

By repeating for all $w \in [W]$, we complete the proof for argument (2). □

Existence of Constrained NE Policy Before we introduce algorithms finding constrained NE(s) in MFG, we first investigate their existence, which is actually directly implied by Prop. 5.1.

Corollary G.3. *Given $\boldsymbol{M}$ satisfying Lipschitz continuity conditions in Assump. D, the MFG M_{MFG} converted from $\boldsymbol{M}$ has at least one constrained NE satisfying $\forall \pi \in \Pi^\dagger,\ J_{M_{MFG}}(\pi, \widehat{\pi}_{Cstr}^{NE}) \leq J_{M_{MFG}}(\widehat{\pi}_{Cstr}^{NE}, \widehat{\pi}_{Cstr}^{NE}) + \varepsilon$ with $\varepsilon = 0$.*

Proof. From Prop. E.12, any MT-MFG $\boldsymbol{M}$ satisfying Assump. D has at least one NE. As implied by Prop. 5.1 when $\varepsilon \to 0$, any MFG M_{MFG} converted from an MT-MFG $\boldsymbol{M}$ with NE(s) should have at least one constrained NE. Therefore, under Assump. D, we can guarantee any model in the converted function class $\mathcal{M}_{\mathrm{MFG}}$ has at least one constrained NE. □

G.5. Algorithm Details

We first generalize some notations in Sec. 4. We define the (constrained) conditional distance between models:

$$d^\dagger(M, \widetilde{M}|\pi) := \max_{\widetilde{\pi}\in\Pi^\dagger} \max\{\mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu_{M,h}^{\pi}) - \mathbb{P}_{\widetilde{M},h}(\cdot|\cdot,\cdot,\mu_{\widetilde{M},h}^{\pi})\|_1],$$
$$\mathbb{E}_{\widetilde{\pi},\widetilde{M}(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu_{M,h}^{\pi}) - \mathbb{P}_{\widetilde{M},h}(\cdot|\cdot,\cdot,\mu_{\widetilde{M},h}^{\pi})\|_1]\}.$$

Besides, given a MF-MDP class $\mathcal{M}$, a model $M \in \mathcal{M}$, and any policy π, we define the ε_0-neighborhood of M in $\mathcal{M}$ w.r.t. distance $d^\dagger(\cdot,\cdot|\pi)$ to be: $\mathcal{B}_\pi^{\dagger,\varepsilon_0}(M;\mathcal{M}) := \{M' \in \mathcal{M} | d^\dagger(M, M'|\pi) \leq \varepsilon_0\}$. The "Central Model" of $\mathcal{M}$ w.r.t. policy π and distance $d^\dagger$ is defined to be the model with the largest neighborhood set $M_{\mathrm{Ctr}}^{\dagger,\varepsilon_0}(\pi;\mathcal{M}) \leftarrow \arg\max_{M\in\mathcal{M}} |\mathcal{B}_\pi^{\dagger,\varepsilon_0}(M;\mathcal{M})|$. When ε_0 and $\mathcal{M}$ is clear from context, we will use $M_{\mathrm{Ctr}}^{\dagger,\pi}$ as a short note.

Besides, we define $\mathcal{E}_M^{\dagger,\mathrm{NE}}(\pi) := \max_{\widetilde{\pi}\in\Pi^\dagger} \Delta_M(\widetilde{\pi},\pi) = \max_{\widetilde{\pi}\in\Pi^\dagger} J_M(\widetilde{\pi},\pi) - J_M(\pi,\pi)$ to be the constrained NE gap.

Algorithm 4: Multi-Type MFG Learning with Constrained Policy Space

Input: Model Class $\boldsymbol{\mathcal{M}}$; Policy Class $\Pi^\dagger$; Accuracy level $\varepsilon_0, \widetilde{\varepsilon}, \bar{\varepsilon}$; Confidence level δ
Convert $\boldsymbol{\mathcal{M}}$ to $\mathcal{M}_{\mathrm{MFG}}$ as described in Appx. G.2; $\mathcal{M}_{\mathrm{MFG}}^1 \leftarrow \mathcal{M}_{\mathrm{MFG}}$, $\delta_0 \leftarrow \frac{\delta}{\log_2|\mathcal{M}_{\mathrm{MFG}}|+1}$.
for $k = 1, 2, \ldots$ **do**
  $\pi^k \leftarrow \arg\min_{\pi\in\Pi^\dagger} |\mathcal{B}_\pi^{\dagger,\varepsilon_0}(M_{\mathrm{Ctr}}^{\dagger,\pi};\mathcal{M}_{\mathrm{MFG}}^k)|$;
  if $|\mathcal{B}_{\pi^k}^{\dagger,\varepsilon_0}(M_{Ctr}^{\dagger,\pi^k};\mathcal{M}_{MFG}^k)| \leq \frac{|\mathcal{M}_{MFG}^k|}{2}$ **then** $\mathcal{M}_{\mathrm{MFG}}^{k+1} \leftarrow$ `ModelElimCstr`$(\pi^k, \mathcal{M}_{\mathrm{MFG}}^k, \widetilde{\varepsilon}, \delta_0)$. ;
  else
    $\pi_{\mathrm{Br}}^{\dagger,\mathrm{NE},k} \leftarrow$ `BridgePolicyCstr`$(\mathcal{M}_{\mathrm{MFG}}^k, \bar{\varepsilon})$;
    $\mathcal{M}_{\mathrm{MFG}}^{k+1} \leftarrow$ `ModelElimCstr`$(\pi_{\mathrm{Br}}^{\dagger,\mathrm{NE},k}, \mathcal{M}_{\mathrm{MFG}}^k, \widetilde{\varepsilon}, \delta_0)$;
    Randomly pick $\widetilde{M}^k$ from $\mathcal{M}_{\mathrm{MFG}}^{k+1}$;
    $\mathcal{E}_{\widetilde{M}^k}^{\dagger,\mathrm{NE}}(\pi_{\mathrm{Br}}^{\dagger,\mathrm{NE},k}) \leftarrow \max_{\pi\in\Pi^\dagger} J_{\widetilde{M}^k}(\pi, \pi_{\mathrm{Br}}^{\dagger,\mathrm{NE},k}) - J_{\widetilde{M}^k}(\pi_{\mathrm{Br}}^{\dagger,\mathrm{NE},k}, \pi_{\mathrm{Br}}^{\dagger,\mathrm{NE},k})$;
    if $\mathcal{E}_{\widetilde{M}^k}^{\dagger,NE}(\pi_{Br}^{\dagger,NE,k}) \leq \frac{3\varepsilon}{4}$ **then return** $\pi_{Br}^{\dagger,NE,k}$;
  end
  if $|\mathcal{M}_{MFG}| = 1$ **then** Return the NE of the model in $\mathcal{M}_{\mathrm{MFG}}$. ;
end
Return the constrained NE policy of the model in $\mathcal{M}_{\mathrm{MFG}}^k$.

Algorithm 5: ModelElimCstr

Input: Reference Policy π; Policy Class $\Pi^\dagger$; Model Class $\bar{\mathcal{M}}$; Accuracy level $\widetilde{\varepsilon}$; Confidence δ
$\bar{\mathcal{M}}^1 \leftarrow \mathcal{M}$; $\bar{\varepsilon}$; Choosing T according to Thm. G.4
for $t = 1, 2, ..., T$ **do**
  $\widetilde{\pi}^t, M^t, M'^t \leftarrow \arg\max_{\widetilde{\pi}\in\Pi^\dagger} \max_{M,M'\in\bar{\mathcal{M}}^t} \mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^H \|\mathbb{P}^w_{M,h}(\cdot|\cdot,\cdot,\mu^\pi_{M,h}) - \mathbb{P}^w_{M',h}(\cdot|\cdot,\cdot,\mu^\pi_{M',h})\|_1]$.
  Denote the value taken at the above as $\Delta^t_{\max}$.
  if $\Delta^t_{\max} \le \widetilde{\varepsilon}$ **then return** $\bar{\mathcal{M}}^t$;
  else
    $\mathcal{Z}^t \leftarrow \{\}$
    for $h = 1, 2..., H$ **do**
      for $w \in [W]$ **do**
        // Trajectory sampling in M^*_{MFG} can be implemented by Def. G.1.
        Sample a trajectory with (π, π), and collect the data at step h: $\{(s^{w,t}_h, a^{w,t}_h, s'^{w,t}_{h+1})\}$.
        Sample a trajectory with $(\widetilde{\pi}^t, \pi)$, and collect the data at step h: $\{(\widetilde{s}^{w,t}_h, \widetilde{a}^{w,t}_h, \widetilde{s}'^{w,t}_{h+1})\}$.
        $\mathcal{Z}^t \leftarrow \mathcal{Z}^t \cup \{(s^{w,t}_h, a^{w,t}_h, s'^{w,t}_{h+1})\}\{(\widetilde{s}^{w,t}_h, \widetilde{a}^{w,t}_h, \widetilde{s}'^{w,t}_{h+1})\}$.
      end
    end
    $\forall M \in \bar{\mathcal{M}}^t$, define

$$l^\pi_{\text{MLE}}(M; \mathcal{Z}^t) := \sum_{k=1}^K \sum_{w=1}^W \sum_{h=1}^H \log \mathbb{P}^w_{M,h}(s'^{w,t}_{h+1}|s^{w,t}_h, a^{w,t}_h, \mu^\pi_{M,h}) + \log \mathbb{P}^w_{M,h}(\widetilde{s}'^{w,t}_{h+1}|\widetilde{s}^{w,t}_h, \widetilde{a}^{w,t}_h, \mu^\pi_{M,h}).$$

    $\bar{\mathcal{M}}^{t+1} \leftarrow \{M \in \bar{\mathcal{M}}^t |\ l^\pi_{\text{MLE}}(M; \mathcal{Z}^t) \ge \max_{\widetilde{M}} l^\pi_{\text{MLE}}(\widetilde{M}; \mathcal{Z}^t) - \log \frac{WHT|\mathcal{M}|}{\delta}\}$.
  end
end

Algorithm 6: BridgePolicyCstr

Input: MF-MDP model class $\mathcal{M}$; Policy Space $\Pi^\dagger$; Accuracy Level $\bar{\varepsilon}$, ε_0
Convert Policy-Aware MDP Model Class $\ddot{\mathcal{M}}$ from $\mathcal{M}$ by Eq. (1).
Construct $\bar{\varepsilon}$-cover of the policy space $\Pi^\dagger$ w.r.t. $d_{\infty,1}$ distance, denoted as $\Pi^\dagger_{\bar{\varepsilon}}$.
for $\widetilde{\pi} \in \Pi^\dagger_{\bar{\varepsilon}}$ **do** Find the central model $\ddot{M}^{\dagger,\varepsilon_0}_{\text{Ctr}}(\widetilde{\pi}; \ddot{\mathcal{M}}) \leftarrow \arg\max_{\ddot{M}\in\ddot{\mathcal{M}}} |\mathcal{B}^{\dagger,\varepsilon_0}_\pi(\ddot{M}; \ddot{\mathcal{M}})|$;
Construct the new PAM $\ddot{M}_{\text{Br}}$ with transition and reward functions $\forall w \in [W], h \in [H]$:

$$\ddot{\mathbb{P}}^w_{\text{Br},h}(\cdot|s_h, a_h, \pi) := \frac{\sum_{\widetilde{\pi}\in\Pi^\dagger_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ \ddot{\mathbb{P}}^w_{\ddot{M}^{\widetilde{\pi}},h}(\cdot|s_h, a_h, \widetilde{\pi})}{\sum_{\widetilde{\pi}\in\Pi^\dagger_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+}$$

$$\ddot{r}^w_{\text{Br},h}(s_h, a_h, \pi) := \frac{\sum_{\widetilde{\pi}\in\Pi^\dagger_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+ \ddot{r}^w_{\ddot{M}^{\widetilde{\pi}},h}(s_h, a_h, \widetilde{\pi})}{\sum_{\widetilde{\pi}\in\Pi^\dagger_{\bar{\varepsilon}}}[2\bar{\varepsilon} - d_{\infty,1}(\pi, \widetilde{\pi})]^+}.$$

Find the NE of bridge model: $\pi^{\dagger,\text{NE}}_{\text{Br}} \leftarrow \arg\min_{\pi\in\Pi^\dagger} \max_{\widetilde{\pi}\in\Pi^\dagger} \ddot{J}_{\ddot{M}_{\text{Br}}}(\widetilde{\pi}; \pi) - \ddot{J}_{\ddot{M}_{\text{Br}}}(\pi; \pi)$.
return $\pi^{\dagger,\text{NE}}_{\text{Br}}$.

G.6. Proofs for Algorithm 4

Theorem G.4. *Under Assump. C and D, in Alg. 5, given any $\widetilde{\varepsilon}$, reference policy π, $\delta \in (0,1)$, and $M^* \in \bar{\mathcal{M}}$, by choosing $T = \widetilde{O}(\frac{H^4}{\widetilde{\varepsilon}^2}(\dim_{\mathrm{CPE}|\Pi^\dagger}(\mathcal{M}, \varepsilon') \wedge (1+\boldsymbol{L}_T)^{2H}(1+\boldsymbol{L}_T H)^2 \dim^{\Pi}_{\mathrm{CPE}|\Pi^\dagger}(\mathcal{M}, \varepsilon')))$ with $\varepsilon' = O(\frac{\widetilde{\varepsilon}}{H^2(1+\boldsymbol{L}_T)^H})$, , w.p. $1-\delta$, the algorithm terminates at some $T_0 \le T$, and return $\bar{\mathcal{M}}^{T_0}$ satisfying (i) $M^* \in \bar{\mathcal{M}}^{T_0}$ (ii) $\forall M \in \bar{\mathcal{M}}^{T_0}$, $d^\dagger(M^*, M|\pi) \le \widetilde{\varepsilon}$.*

Proof. The proof is the same as Thm. F.3, except that we consider the constrained policy space, and need to replace P-MBED with constrained P-MBED. □

Theorem G.5. *Suppose we feed Alg. 6 with a model class $\ddot{\mathcal{M}}$ and policy space $\Pi^\dagger$, then for the bridge model $\ddot{M}_{Br}$ it computes, by choosing $\bar{\varepsilon} = \varepsilon_0 / \min\{2HL_r \frac{(1+\boldsymbol{L}_T)^H - 1}{\boldsymbol{L}_T}, 2H(H+1)((1+\boldsymbol{L}_T)^H - 1)\}$, for any reference policy $\pi \in \Pi^\dagger$ and its associated central model $\ddot{M}^{\dagger,\varepsilon_0}_{Ctr}(\pi; \ddot{\mathcal{M}})$, we have:*

$$\max_{\widetilde{\pi} \in \Pi^\dagger} \mathbb{E}_{\widetilde{\pi}, \ddot{M}^{\dagger,\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}})(\pi)}[\sum_{h=1}^H \|\ddot{\mathbb{P}}_{\ddot{M}^{\dagger,\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}}),h}(\cdot|s_h, a_h, \pi) - \ddot{\mathbb{P}}_{Br,h}(\cdot|s_h, a_h, \pi)\|_1] \le (H+3)\varepsilon_0,$$

$$\max_{\widetilde{\pi} \Pi^\dagger} \mathbb{E}_{\widetilde{\pi}, \ddot{M}^{\dagger,\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}})(\pi)}[\sum_{h=1}^H |r_{\ddot{M}^{\dagger,\varepsilon_0}_{Ctr}(\pi;\ddot{\mathcal{M}}),h}(s_h, a_h, \pi) - \ddot{r}_{Br,h}(s_h, a_h, \pi)|] \le L_r H(H+4)\varepsilon_0.$$

Proof. The proof is the same as Thm. F.5 except that we constrain the policies in $\Pi^\dagger$. □

Lemma G.6. *Suppose the `Else`-branch in Line 6 if activated in Alg. 2, for policy $\pi^{\dagger,NE,k}_{Br}$ and its corresponding central model $M^{\dagger,k}_{Ctr} := \arg\max_{M \in \mathcal{M}^k} |\mathcal{B}^{\dagger,\varepsilon_0}_{\pi^{\dagger,NE,k}_{Br}}(M; \mathcal{M}^k)|$, we have:*

$$\mathcal{E}^{\dagger,NE}_{M^{\dagger,k}_{Ctr}}(\pi^{\dagger,NE,k}_{Br}) := \max_{\pi \in \Pi^\dagger} \Delta_{M^{\dagger,k}_{Ctr}}(\pi, \pi^{\dagger,NE,k}_{Br}) \le 2(1+L_r)(H+4)\varepsilon_0.$$

Proof. The proof is the almost the same as Lem. F.6, except that we consider the constrained policy space.

For any policy $\pi \in \Pi^\dagger$, we have

$$\begin{aligned}
&\Delta_{M^{\dagger,k}_{\mathrm{Ctr}}}(\pi, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) \\
\le& \Delta_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}}}(\pi, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) - \Delta_{\ddot{M}_{\mathrm{Br}}}(\pi, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) && (\Delta_{\ddot{M}_{\mathrm{Br}}}(\pi, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) \le 0) \\
\le& |\ddot{J}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}}}(\pi, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) - \ddot{J}_{\ddot{M}_{\mathrm{Br}}}(\pi, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})| + |\ddot{J}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}}}(\pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) - \ddot{J}_{\ddot{M}_{\mathrm{Br}}}(\pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})| \\
\le& \mathbb{E}_{\pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}, \ddot{M}^{\dagger,k}_{\mathrm{Ctr}}(\pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})}[\sum_{h=1}^H |\ddot{r}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}},h}(s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) - \ddot{r}_{\ddot{M}_{\mathrm{Br}},h}(s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})| \\
&\qquad + \|\ddot{\mathbb{P}}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}},h}(\cdot|s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}), \ddot{\mathbb{P}}_{\ddot{M}_{\mathrm{Br}},h}(\cdot|s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})\|_1] \\
&+ \mathbb{E}_{\pi, \ddot{M}^{\dagger,k}_{\mathrm{Ctr}}(\pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})}[\sum_{h=1}^H |\ddot{r}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}},h}(s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) - \ddot{r}_{\ddot{M}_{\mathrm{Br}},h}(s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})| \\
&\qquad + \|\ddot{\mathbb{P}}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}},h}(\cdot|s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}), \ddot{\mathbb{P}}_{\ddot{M}_{\mathrm{Br}},h}(\cdot|s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})\|_1] \\
\le& 2\max_{\pi \in \Pi^\dagger} \mathbb{E}_{\pi, \ddot{M}^{\dagger,k}_{\mathrm{Ctr}}(\pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})}[\sum_{h=1}^H |\ddot{r}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}},h}(s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}) - \ddot{r}_{\ddot{M}_{\mathrm{Br}},h}(s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})|] \\
&+ 2\max_{\pi \in \Pi^\dagger} \mathbb{E}_{\pi, \ddot{M}^{\dagger,k}_{\mathrm{Ctr}}(\pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})}[\sum_{h=1}^H \|\ddot{\mathbb{P}}_{\ddot{M}^{\dagger,k}_{\mathrm{Ctr}},h}(\cdot|s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}}), \ddot{\mathbb{P}}_{\ddot{M}_{\mathrm{Br}},h}(\cdot|s_h, a_h, \pi^{\dagger,\mathrm{NE},k}_{\mathrm{Br}})\|_1] \\
\le& 2(1+L_r H)(H+4)\varepsilon_0. && \text{(Thm. G.5)}
\end{aligned}$$

which finishes the proof. □

Theorem G.7. *In Alg. 4, by choosing $\varepsilon_0 = \frac{\varepsilon}{8(H+4)(1+L_rH)}$, $\widetilde{\varepsilon} = \frac{\varepsilon_0}{6}$ and choosing $\bar{\varepsilon}$ according to Thm. G.5, on the good events in Thm. G.4, (1) if the* `If-Branch` *in Line 5 is activated: we have $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$; (2) otherwise, in the* `Else-Branch` *in Line 6: either we return the $\pi_{Br}^{\dagger,NE,k}$ which is an ε-approximate NE for M^*; or the algorithm continues with $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k|/2$.*

Proof. We separately discuss the if and else branches in the algorithm.

Proof for `If-Branch` in Line 5 On the events in Thm. G.4, for any $\widetilde{M} \notin \mathcal{B}^{\dagger,\varepsilon_0}_{\pi^k}(M^*;\mathcal{M}^k)$, we have $d^\dagger(M^*,\widetilde{M}) \geq \varepsilon_0 > \widetilde{\varepsilon}$, which implies $\widetilde{M} \notin \mathcal{M}^{k+1}$. Combining the condition of `If-Branch`, we have:

$$|\mathcal{M}^{k+1}| \leq |\mathcal{B}^{\dagger,\varepsilon_0}_{\pi^k}(M^*;\mathcal{M}^k)| \leq \frac{|\mathcal{M}^k|}{2}.$$

Proof for `Else-Branch` in Line 6 First of all, on the events in Thm. G.4, we have $d^\dagger(M^*,\widetilde{M}^k|\pi_{\text{Br}}^{\dagger,\text{NE},k}) \leq \widetilde{\varepsilon}$. By applying Lem. I.2, it implies:

$$\begin{aligned}
&|\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) - \Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k})| \\
\leq&\mathbb{E}_{\pi,M^*(\pi_{\text{Br}}^{\dagger,\text{NE},k})}[\sum_{h=1}^H \|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu_{M^*,h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}}) - \mathbb{P}_{\widetilde{M}^k,h}(\cdot|s_h,a_h,\mu_{\widetilde{M}^k,h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}})\|_1] \\
&+(2L_rH+1)\mathbb{E}_{\pi_{\text{Br}}^{\dagger,\text{NE},k},M^*(\pi_{\text{Br}}^{\dagger,\text{NE},k})}[\sum_{h=1}^H \|\mathbb{P}_{M^*,h}(\cdot|s_h,a_h,\mu_{M^*,h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}}) - \mathbb{P}_{\widetilde{M}^k,h}(\cdot|s_h,a_h,\mu_{\widetilde{M}^k,h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}})\|_1] \\
\leq&2(L_rH+1)\widetilde{\varepsilon}.
\end{aligned} \tag{17}$$

Also note that:

$$\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) = \Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) - \Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) + \Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}).$$

In the following, we separately discuss two cases.

Case 1: $\mathcal{E}^{\dagger,\text{NE}}_{\widetilde{M}^k}(\pi_{\text{Br}}^{\dagger,\text{NE},k}) \leq \frac{3\varepsilon}{4}$ and Line 9 is activated Given that $\widetilde{\varepsilon} \leq \frac{\varepsilon}{16(1+L_rH)}$:

$$\begin{aligned}
\forall \pi \in \Pi^\dagger, \quad \Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) \leq& |\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) - \Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k})| + \mathcal{E}^{\dagger,\text{NE}}_{\widetilde{M}^k}(\pi_{\text{Br}}^{\dagger,\text{NE},k}) \\
&(\mathcal{E}^{\dagger,\text{NE}}_{\widetilde{M}^k}(\pi_{\text{Br}}^{\dagger,\text{NE},k}) = \max_{\pi\in\Pi^\dagger}\Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k})) \\
\leq& 2(L_rH+1)\widetilde{\varepsilon} + \frac{3\varepsilon}{4} \leq \varepsilon.
\end{aligned}$$

which implies $\pi_{\text{Br}}^{\dagger,\text{NE},k}$ is an ε-NE of M^*.

Case 2: $\mathcal{E}^{\dagger,\text{NE}}_{\widetilde{M}^k}(\pi_{\text{Br}}^{\dagger,\text{NE},k}) > \frac{3\varepsilon}{4}$ and Line 9 is not activated As a result, for any policy $\pi \in \Pi^\dagger$, by Eq. (17), we have:

$$\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) \geq -|\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) - \Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k})| + \Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) \geq \Delta_{\widetilde{M}^k}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) - 2(L_rH+1)\widetilde{\varepsilon}.$$

Therefore, by our choice of $\widetilde{\varepsilon}$,

$$\max_{\pi\in\Pi^\dagger}\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) \geq \mathcal{E}^{\dagger,\text{NE}}_{\widetilde{M}^k}(\pi_{\text{Br}}^{\dagger,\text{NE},k}) - 2(L_rH+1)\widetilde{\varepsilon} \geq \frac{5\varepsilon}{8}.$$

On the other hand, by Lem. F.6, for any $\pi \in \Pi^\dagger$, we have:

$$\begin{aligned}
&\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k}) - 2(1+L_rH)(H+4)\varepsilon_0 \\
\leq&|\Delta_{M^*}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k})| - |\Delta_{M_{\text{Cr}}^{\dagger,k}}(\pi,\pi_{\text{Br}}^{\dagger,\text{NE},k})| \qquad \text{(Here we apply Lem. G.6)}
\end{aligned}$$

$$
\begin{aligned}
&\leq |\Delta_{M^*}(\pi, \pi_{\text{Br}}^{\dagger,\text{NE},k}) - \Delta_{M_{\text{Ctr}}^{\dagger,k}}(\pi, \pi_{\text{Br}}^{\dagger,\text{NE},k})| \\
&\leq \mathbb{E}_{\pi, M^*(\pi_{\text{Br}}^{\dagger,\text{NE},k})}[\sum_{h=1}^{H} \|\mathbb{P}_{M^*,h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}}) - \mathbb{P}_{M_{\text{Ctr}}^{\dagger,k},h}(\cdot|s_h, a_h, \mu_{\widetilde{M},h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}})\|_1] \\
&\quad + (2L_rH+1)\mathbb{E}_{\pi_{\text{Br}}^{\dagger,\text{NE},k}, M^*(\pi_{\text{Br}}^{\dagger,\text{NE},k})}[\sum_{h=1}^{H} \|\mathbb{P}_{M^*,h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}}) - \mathbb{P}_{M_{\text{Ctr}}^{\dagger,k},h}(\cdot|s_h, a_h, \mu_{M_{\text{Ctr}}^{\dagger,k},h}^{\pi_{\text{Br}}^{\dagger,\text{NE},k}})\|_1] \\
&\leq (2L_rH+2)d^\dagger(M^*, M_{\text{Ctr}}^{\dagger,k}|\pi_{\text{Br}}^{\dagger,\text{NE},k}).
\end{aligned}
$$

According to the choice of ε_0, we have $2(1+L_rH)(H+4)\varepsilon_0 \leq \frac{\varepsilon}{4}$, therefore,

$$
d^\dagger(M^*, M_{\text{Ctr}}^{\dagger,k}|\pi_{\text{Br}}^{\dagger,\text{NE},k}) \geq \frac{1}{2(L_rH+1)}\Big(\max_{\pi\in\Pi^\dagger} |\Delta_{M^*}(\pi, \pi_{\text{Br}}^{\dagger,\text{NE},k})| - 2(1+L_rH)(H+4)\varepsilon_0\Big) \geq \frac{3\varepsilon}{16(L_rH+1)}.
$$

Next we try to show that models in $\mathcal{B}^{\varepsilon_0}_{\pi_{\text{Br}}^{\dagger,\text{NE},k}}(M_{\text{Ctr}}^{\dagger,k}, \mathcal{M}^k)$ will be eliminated. For any $M \in \mathcal{B}^{\varepsilon_0}_{\pi_{\text{Br}}^{\dagger,\text{NE},k}}(M_{\text{Ctr}}^{\dagger,k}, \mathcal{M}^k)$, we have:

$$
d^\dagger(M, M^*|\pi_{\text{Br}}^{\dagger,\text{NE},k}) \geq d^\dagger(M_{\text{Ctr}}^{\dagger,k}, M^*|\pi_{\text{Br}}^{\dagger,\text{NE},k}) - d^\dagger(M, M_{\text{Ctr}}^{\dagger,k}|\pi_{\text{Br}}^{\dagger,\text{NE},k}) \geq \frac{3\varepsilon}{16(L_rH+1)} - \varepsilon > \widetilde{\varepsilon}.
$$

On the good events in Thm. G.4, we have $M \notin \mathcal{M}^{k+1}$, which implies

$$
|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k| - |\mathcal{B}^{\varepsilon_0}_{\pi_{\text{Br}}^{\dagger,\text{NE},k}}(M_{\text{Ctr}}^{\dagger,k}, \mathcal{M}^k)| \leq |\mathcal{M}^k|/2.
$$

□

Theorem G.8. *[Sample Complexity in MT-MFG] Under Assump. C and D, by running Alg. 4 with Alg. 5 as* `ModelElimCstr` *and Alg. 6 as* `BridgePolicyCstr`*, and hyper-parameter choices according to Thm. G.4, G.5, and G.7, w.p.* $1-\delta$*, Alg. 4 will terminate at some* $k \leq \log_2|\boldsymbol{\mathcal{M}}|+1$ *and return an* ε*-NE of* $\boldsymbol{M}^*$*. The number of trajectories consumed is* $\widetilde{O}(\frac{W^2H^7}{\varepsilon^2}(1+\boldsymbol{L}_rH)^2\sum_{w\in[W]}(\dim_{\text{MTPE}}(\mathcal{M}^w,\varepsilon') \wedge H^3(1+\boldsymbol{L}_T)^{2H}(1+\boldsymbol{L}_TH)^2\dim^{\text{II}}_{\text{MTPE}}(\mathcal{M}^w,\varepsilon'))\log^2\frac{|\boldsymbol{\mathcal{M}}|}{\delta})$*, where* $\varepsilon' = O(\varepsilon/WH^3(1+\boldsymbol{L}_rH)(1+\boldsymbol{L}_T)^H)$*,* $\dim_{\text{MTPE}}(\mathcal{M}^w,\varepsilon')$ *and* $\dim^{\text{II}}_{\text{MTPE}}(\mathcal{M}^w,\varepsilon')$ *are the Multi-Type P-MBED defined in Def. D.10, and we omit the logarithmic terms of* $H, \varepsilon, \log|\boldsymbol{\mathcal{M}}|, \dim_{\text{MTPE}}, 1+\boldsymbol{L}_T$ *and* $1+\boldsymbol{L}_r$*.*

Proof. As a result of Thm. G.7, w.p. $1-\frac{\delta}{\log_2|\mathcal{M}_{\text{MFG}}|+1}\cdot(\log_2|\mathcal{M}_{\text{MFG}}|+1) = 1-\delta$, there exists a step $k \leq \log_2|\mathcal{M}_{\text{MFG}}|+1 = \log_2|\boldsymbol{\mathcal{M}}|+1$ such that Alg. 4 will terminate the return us an $\frac{\varepsilon}{W}$-approximate NE of M^*_{MFG}. The total number of trajectories required is:

$$
(\log_2|\boldsymbol{\mathcal{M}}|+1)\cdot T\cdot 2H = \widetilde{O}(\frac{H^5}{\widetilde{\varepsilon}^2}\Big(\dim_{\text{CPE}|\Pi^\dagger}(\mathcal{M}_{\text{MFG}},\varepsilon') \wedge H^3(1+\boldsymbol{L}_T)^{2H}(1+\boldsymbol{L}_TH)^2\dim^{\text{II}}_{\text{CPE}|\Pi^\dagger}(\mathcal{M}_{\text{MFG}},\varepsilon')\Big))
$$

Note that in Thm. G.4, we choose $\widetilde{\varepsilon} = \frac{\varepsilon_0}{6} = O(\frac{\varepsilon}{WH(1+L_rH)})$, and $\varepsilon' = O(\widetilde{\varepsilon}/H^2(1+L_T)^H) = O(\varepsilon/WH^3(1+L_rH)(1+L_T)^H)$. Combining with the above discussion and Prop. D.14 and Prop. 5.1, we finish the proof. □

H. Approximation Ability of Multi-Type MFGs

H.1. Multi-Type Symmetric Anonymous Games

Notations Given a multi-agent system where agents are divided into W groups, where for each type w the agents share the state-action spaces $\mathcal{S}^w, \mathcal{A}^w$ and initial distribution μ_1^w, we use N^w to denote the number of agents in group $w \in [W]$, and $s_h^{w,n}, a_h^{w,n}$ and $\pi^{w,n}$ to denote the state, action, and policy for the n-th agent in type w, respectively. Besides, we define $\boldsymbol{s}_h := \{s_h^{w,n}\}_{w\in[W],n\in[N^w]}, \boldsymbol{a}_h := \{a_h^{w,n}\}_{w\in[W],n\in[N^w]}$ to be the collection of states and actions of all agents in the system at step h, and denote $p_{\boldsymbol{s}_h} := \{p_{\boldsymbol{s}_h}^1, ..., p_{\boldsymbol{s}_h}^W\}$ to be the empirical distribution of the agents' states with:

$$p_{\boldsymbol{s}_h}^w \in \mathbb{R}^{|\mathcal{S}^w|}: \quad p_{\boldsymbol{s}_h}^w(\cdot) = \frac{1}{N^w}\sum_{n=1}^{N^w}\delta(s_h^{w,n} = \cdot),$$

To distinguish the policy in MFG setting, we use $\widetilde{\boldsymbol{\nu}} := \{\pi^{w,n}\}_{w\in[W],n\in[N^w]}$ to denote the collection of policies. We will denote $\boldsymbol{\nu}(\boldsymbol{a}_h|\boldsymbol{s}_h) := \prod_{w\in[W]}\prod_{n\in[N^w]}\pi^{w,n}(a_h^{w,n}|s_h^{w,n})$. Besides, we use $\boldsymbol{\nu}^{-(w,n)} \circ \widetilde{\pi}^{w,n}$ to denote the policy replacing $\pi^{w,n}$ to $\widetilde{\pi}^{w,n}$ while keeping the others fixed.

Definition H.1 (Multi-Type Symmetric Anonymous Game)**.** The Multi-Type Symmetric Anonymous Game (MT-SAG) $\bar{\boldsymbol{M}} := \{(\mu_1^w, \mathcal{S}^w, \mathcal{A}^w, H, \mathbb{P}^w, r^w)\}_{w\in[H]}$ is a Multi-Agent system consists of W groups. Given a policy $\boldsymbol{\pi}$, the system evolves as:

$$s_1^{w,n} \sim \mu_1^w;\ \forall h, w, n: \quad a_h^{w,n} \sim \pi_h^{w,n}(\cdot|s_h^{w,n}),\ r_h^{w,n} \leftarrow r_h^w(s_h^{w,n}, a_h^{w,n}, p_{\boldsymbol{s}_h}),\ s_{h+1}^{w,n} \sim \mathbb{P}_h^w(\cdot|s_h^w, a_h^w, p_{\boldsymbol{s}_h}). \tag{18}$$

Given a policy $\boldsymbol{\nu}$, we define the value functions $V: \mathcal{S} \to [0,1]$ and $Q: \mathcal{S}\times\mathcal{A} \to [0,1]$ of the (w,n)-th agent conditioning on the system state $\boldsymbol{s}_h$ to be:

$$V_{\bar{\boldsymbol{M}},h}^{w,n,\boldsymbol{\nu}}(s_h^{w,n}; \boldsymbol{s}_h) := \mathbb{E}_{\bar{\boldsymbol{M}},\boldsymbol{\nu}}[\sum_{h'=h}^{H} r_{h'}^w(s_{h'}^{w,n}, a_{h'}^{w,n}, p_{\boldsymbol{s}_{h'}})|\boldsymbol{s}_h]; \tag{19}$$

where the expectation is taken over the evolution process in Eq. (18). Besides, we define the total value starting from the initial states $J_{\bar{\boldsymbol{M}}}^{w,n}(\boldsymbol{\nu}) := \mathbb{E}_{\boldsymbol{\mu}_1}[V_1^{w,n,\boldsymbol{\nu}}(s_1^{w,n}; \boldsymbol{s}_1)]$.

A policy $\boldsymbol{\nu}$ is called to be the NE policy in MT-SAG if any agent can not improve its value by deviating from its current policy while the others' are fixed,

$$\forall w, n, \quad \max_{\widetilde{\pi}^{w,n}} J_{\bar{\boldsymbol{M}}}^{w,n}(\boldsymbol{\nu}^{-(w,n)} \circ \widetilde{\pi}^{w,n}) \le J_{\bar{\boldsymbol{M}}}^{w,n}(\boldsymbol{\nu}).$$

and a policy $\boldsymbol{\nu}'$ is called to be an ε-approximate NE in MT-SAG if

$$\forall w, n, \quad \max_{\widetilde{\pi}^{w,n}} J_{\bar{\boldsymbol{M}}}^{w,n}(\boldsymbol{\nu}^{-(w,n)} \circ \widetilde{\pi}^{w,n}) \le J_{\bar{\boldsymbol{M}}}^{w,n}(\boldsymbol{\nu}) + \varepsilon.$$

Assumption E (Lipschitz Continuity in MT-SAG)**.** We assume the transition and reward functions of MT-SAG are Lipschitz continuous w.r.t. the density, s.t. $\forall w \in [W], h \in [H], \quad \forall \widehat{\boldsymbol{\mu}}_h, \widehat{\boldsymbol{\mu}}_h' \in \Delta(\mathcal{S}^1)\times ...\Delta(\mathcal{S}^W)$

$$\|\mathbb{P}_h^w(\cdot|s_h^w, a_h^w, \widehat{\boldsymbol{\mu}}_h) - \mathbb{P}_h^w(\cdot|s_h^w, a_h^w, \widehat{\boldsymbol{\mu}}_h')\|_1 \le \frac{\boldsymbol{L}_T}{W}\|\widehat{\boldsymbol{\mu}}_h - \widehat{\boldsymbol{\mu}}_h'\|_1$$
$$|r_h^w(s_h^w, a_h^w, \widehat{\boldsymbol{\mu}}_h) - r_h^w(s_h^w, a_h^w, \widehat{\boldsymbol{\mu}}_h')\|_1 \le \frac{\boldsymbol{L}_r}{W}\|\widehat{\boldsymbol{\mu}}_h - \widehat{\boldsymbol{\mu}}_h'\|_1.$$

H.2. Approximating MT-SAGs via MT-MFGs

Definition H.2 (Multi-Type Mean-Field Game Approximation of MT-SAG)**.** Given an MT-SAG $\bar{\boldsymbol{M}}$, its Multi-Type Mean-Field (MT-MFG) Approximation is a model Multi-Type MF-MDP model $\boldsymbol{M} := \{(\mu_1^w, \mathcal{S}^w, \mathcal{A}^w, H, \mathbb{P}_{\boldsymbol{M}}^w, r_{\boldsymbol{M}}^w)\}_{w\in[W]}$, sharing the group, initial distribution, state-action spaces and transition $\mathbb{P}^w$ and reward function r^w as MT-SAG (i.e. $\mathbb{P}_{\boldsymbol{M}}^w(\cdot|\cdot,\cdot,\cdot) = \mathbb{P}_{\bar{\boldsymbol{M}}}^w(\cdot|\cdot,\cdot,\cdot), r_{\boldsymbol{M}}^w(\cdot,\cdot,\cdot) = r_{\bar{\boldsymbol{M}}}^w(\cdot,\cdot,\cdot)$), by have different transition rules.

Next, we describe "the different transition rules" in MT-MFG. For simplicity of notation, in the following, we omit $\boldsymbol{M}$ or $\bar{\boldsymbol{M}}$ in the sub-scription of transition and reward functions. Given a reference policy $\boldsymbol{\pi} := \{\pi^1, ..., \pi^W\}$ consisting of W policies shared by each group, the density $\boldsymbol{\mu}_h^{\boldsymbol{\pi}} := \{\mu_h^{1,\boldsymbol{\pi}}, ..., \mu_h^{W,\boldsymbol{\pi}}\}$ at step h is defined by:

$$\mu_1^{w,\boldsymbol{\pi}} = \mu_1^w, \quad \forall h \geq 1, \quad \mu_{h+1}^{w,\boldsymbol{\pi}} \leftarrow \Gamma_h^{w,\boldsymbol{\pi}}(\boldsymbol{\mu}_h^{\boldsymbol{\pi}}), \text{ with } \Gamma_h^{w,\boldsymbol{\pi}}(\boldsymbol{\mu}_h)(\cdot) := \sum_{s_h^w, a_h^w} \mu_h^w(s_h^w)\pi_h^w(a_h^w|s_h^w)\mathbb{P}_h^w(\cdot|s_h^w, a_h^w, \boldsymbol{\mu}_h).$$

where $\Gamma_h^{w,\boldsymbol{\pi}}$ is a mapping from $\Delta(\mathcal{S}^1) \times ...\Delta(\mathcal{S}^W)$ to $\Delta(\mathcal{S}^w)$. The evolution process of the (w, n) agent in type w following a deviation policy $\widetilde{\pi}^w$ conditioning on reference policy $\boldsymbol{\pi}$ is specified by:

$$s_1^{w,n} \sim \mu_1^w;\ \forall h: \quad a_h^{w,n} \sim \widetilde{\pi}_h^w(\cdot|s_h^{w,n}),\ r_h^{w,n} \leftarrow r_h^w(s_h^{w,n}, a_h^{w,n}, \boldsymbol{\mu}_h^{\boldsymbol{\pi}}),\ s_{h+1}^{w,n} \sim \mathbb{P}_h^w(\cdot|s_h^w, a_h^w, \boldsymbol{\mu}_h^{\boldsymbol{\pi}}). \tag{20}$$

Comparing with Eq. (18), the evolution of agents' states is depend on the density when $\forall w \in [W],\ N^w \to +\infty$, instead of the empirical one in practice.

Recall that given a reference policy $\boldsymbol{\pi}$ and a deviation policy $\widetilde{\pi}^w$, the value functions $V : \mathcal{S} \to [0, 1]$ of the (w, n)-th agent conditioning on the density $\boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}}$ is defined to be:

$$V_{\boldsymbol{M},h}^{w,n,\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}(s_h^{w,n}; \boldsymbol{\mu}_{\boldsymbol{M},h}^{\boldsymbol{\pi}}) := \mathbb{E}_{\widetilde{\pi}^w, \boldsymbol{M}(\boldsymbol{\pi})}[\sum_{h'=h}^{H} r_{h'}^w(s_{h'}^{w,n}, a_{h'}^{w,n}, \boldsymbol{\mu}_{\boldsymbol{M},h'}^{\boldsymbol{\pi}})];$$

where the expectation is taken over the process in Eq. (20). Then, we define the total value of $\boldsymbol{\pi}^{-w} \circ \widetilde{\pi}^w$ given the reference policy $\boldsymbol{\pi}$ to be:

$$J_{\boldsymbol{M}}^{w,n}(\boldsymbol{\pi}^{-w} \circ \widetilde{\pi}^w; \boldsymbol{\pi}) := \mathbb{E}_{s_1^w \sim \mu_1^w}[V_1^{w,n,\widetilde{\pi}^w}(s_1^w, \boldsymbol{\mu}_1)].$$

A $\boldsymbol{\pi}$ is called NE policy if:

$$\forall w, n, \forall \widetilde{\pi}^w, \quad J_{\boldsymbol{M}}^{w,n}(\boldsymbol{\pi}^{-w} \circ \widetilde{\pi}^w; \boldsymbol{\pi}) \leq J_{\boldsymbol{M}}^{w,n}(\boldsymbol{\pi}; \boldsymbol{\pi}).$$

Proposition H.3 (Approximation Error of MT-MFG)**.** *Given a Multi-Type Symmetric Anonymous Game (MT-SAG) $\bar{\boldsymbol{M}}$, as defined in Def. H.1, and its Multi-Type MFG approximation (MT-MFG) $\boldsymbol{M}$, as defined in Def. H.2, suppose $\boldsymbol{\pi} := \{\pi^w\}_{w\in[W]}$ is the NE policy of MT-MFG, then for any $\varepsilon_0 > 0$, the lifted policy $\boldsymbol{\nu} := \{\pi^{w,n}\}_{w\in[W],n\in[N^w]}$ with $\pi^{w,n} = \pi^w,\ \forall n \in [N^w]$ is an ε_0-approximate NE of MT-SAG if*

$$\forall w \in [W], \quad N^w \geq O((\boldsymbol{L}_r + \boldsymbol{L}_T)^2 W^2 H^3 \Big(\frac{(1+\boldsymbol{L}_T)^H - 1}{\boldsymbol{L}_T}\Big)^2 \frac{\sum_{w\in[W]} S^w}{\varepsilon_0^2} \log \frac{2(1+\boldsymbol{L}_r + \boldsymbol{L}_T)HWS_{\max}}{\varepsilon_0}).$$

where $S_{\max} := \max_w S^w$.

Proof. Given $\boldsymbol{\pi} := \{\pi^1, ..., \pi^W\}$, we denote $\boldsymbol{\nu} := \{\pi^{w,n}\}_{w\in[W],n\in[N^w]}$ to be the lifted policy such that $\nu^{w,n} \leftarrow \pi^w$ for all w and $n \in [N^w]$. Given a deviation policy $\widetilde{\pi}^w$ for some w ($\widetilde{\pi}^w$ may equal π^w), we define $\widetilde{\boldsymbol{\nu}} := \{\pi^{w,n}\}_{w\in[W],n\in[N^w]}$ to be a policy in MT-SAG, such that $\widetilde{\pi}^{\widetilde{w},n} \leftarrow \pi^{\widetilde{w}}$ for all agent except the $(w, 1)$-th agent (i.e. the first agent in type w), we set $\widetilde{\pi}^{(w,1)} \leftarrow \widetilde{\pi}^w$.

Concentration Events We first provide a high-probability bound for the distance between state density $\boldsymbol{\mu}_h^{\boldsymbol{\pi}}$ in MT-MFG and the empirical distribution $p_{\boldsymbol{s}_h}^{\widetilde{\boldsymbol{\nu}}}$ in MT-SAG w.r.t. the lifted policy $\widetilde{\boldsymbol{\nu}}$.

We use $\Gamma_{\boldsymbol{M},h}^{w,\boldsymbol{\pi}}(\cdot)$ to denote the operator $\Gamma_h^{w,\boldsymbol{\pi}}(\cdot)$ in Eq. (20) specified in model $\boldsymbol{M}$. We extend its definition to $h = 0$ by $\forall \boldsymbol{\mu}_0, \Gamma_0^{w,\widetilde{\boldsymbol{\nu}}}(\boldsymbol{\mu}_0) \leftarrow \mu_1^w$, and define $\Gamma_{\boldsymbol{M},h}^{\boldsymbol{\pi}}(\cdot) := \{\Gamma_{\boldsymbol{M},h}^{1,\boldsymbol{\pi}}(\cdot), ..., \Gamma_{\boldsymbol{M},h}^{W,\boldsymbol{\pi}}(\cdot)\}$. Then, conditioning on $p_{\boldsymbol{s}_{h-1}}^{\widetilde{\boldsymbol{\nu}}}$, we have:

$$\begin{aligned}
&\|\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[p_{\boldsymbol{s}_h}^{\widetilde{\boldsymbol{\nu}}}|p_{\boldsymbol{s}_{h-1}}^{\widetilde{\boldsymbol{\nu}}}] - \Gamma_{\boldsymbol{M},h-1}^{\boldsymbol{\pi}}(p_{\boldsymbol{s}_{h-1}}^{\widetilde{\boldsymbol{\nu}}})\|_1 \\
=&\|\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[p_{\boldsymbol{s}_h}^{w,\widetilde{\boldsymbol{\nu}}}|p_{\boldsymbol{s}_{h-1}}^{\widetilde{\boldsymbol{\nu}}}] - \Gamma_{\boldsymbol{M},h-1}^{w,\boldsymbol{\pi}}(p_{\boldsymbol{s}_{h-1}}^{\widetilde{\boldsymbol{\nu}}})\|_1 \\
=&\frac{N^w - 1}{N^w}\|\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\frac{1}{N^w}\sum_{n=2}^{N^w-1} \boldsymbol{\delta}_{s_h^{w,n}=(\cdot)}|p_{\boldsymbol{s}_{h-1}}^{\widetilde{\boldsymbol{\nu}}}] - \Gamma_{\boldsymbol{M},h-1}^{w,\boldsymbol{\pi}}(p_{\boldsymbol{s}_{h-1}}^{\widetilde{\boldsymbol{\nu}}})\|_1
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{N^w}\|\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\boldsymbol{\delta}_{s_h^{w,1}=(\cdot)}|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}]-\Gamma^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\|_1\\
=&\frac{1}{N^w}\|\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\boldsymbol{\delta}_{s_h^{w,1}=(\cdot)}|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}]-\Gamma^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\|_1\\
\leq&\frac{2}{N^w}.
\end{aligned}
\tag{21}
$$

where $\boldsymbol{\delta}_{s_h^{w,n}=(\cdot)}\in\mathbb{R}^{\sum_w|\mathcal{S}^w|}$ denotes a vector with 1 at the (w,n)-th value and 0 at the others. In the equalities, we use the fact that $\bar{\boldsymbol{M}}$ and $\boldsymbol{M}$ share the transition function.

Besides, conditioning on $p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}$, for any $\widetilde{w}\in[W]$ we can treat $\{s_h^{\widetilde{w},n}\}_{n\in[N^{\widetilde{w}}]}$ as i.i.d. samples according to distribution $\mathbb{E}[p^{\widetilde{w},\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}]$. By applying Lem. I.6 for a fixed $p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}$, $\forall w\in[W],h\in[H]$, for any $\varepsilon\in(0,1)$ and $\delta\in(0,1)$, as long as $N^{\widetilde{w}}\geq\min\{\frac{8W^2\mathcal{S}^{\widetilde{w}}}{\varepsilon^2},\frac{8W^2}{\varepsilon^2}\log\frac{2W}{\delta}\}$ holds for any $\widetilde{w}\in[W]$, we have:

$$
\begin{aligned}
&\Pr(\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}-\Gamma^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\|_1\geq\varepsilon)\\
\leq&\Pr(\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}-\Gamma^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\|+\|\mathbb{E}[p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}]-\Gamma^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\|_1\geq\varepsilon)\\
\leq&\Pr(\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}-\mathbb{E}[p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}]\|_1\geq\frac{\varepsilon}{2})\qquad\text{(Eq. (21))}\\
\leq&\sum_{\widetilde{w}\in[W]}\Pr(\|p^{\widetilde{w},\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}-\mathbb{E}[p^{\widetilde{w},\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}]\|_1\geq\varepsilon/2W)\\
\leq&\delta.
\end{aligned}
\tag{22}
$$

Note that the number of possible values of $p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}$ can be upper bounded by $\prod_{w\in[W]}(N^w)^{S^w}$. We define event $\mathcal{E}:=\{\forall h\in[H],w\in[W],\ \|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}-\Gamma^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\|_1<\varepsilon\}$. By applying a union bound over h,w and all possible $p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}$, we have:

$$
\Pr(\mathcal{E})\geq1-\delta,\quad\text{as long as }\forall w\in[W],\ N^w\geq O(\frac{W^2\sum_{w\in[W]}S^w}{\varepsilon^2}\log\frac{2HWS_{\max}}{\delta\varepsilon}),
\tag{23}
$$

Density Error Decomposition The following discussion are based on the event $\mathcal{E}$. Recall we use $\boldsymbol{\mu}^{\boldsymbol{\pi}}_h:=\{\mu^{1,\boldsymbol{\pi}}_h,...,\mu^{W,\boldsymbol{\pi}}_h\}$ to denote the density induced by $\boldsymbol{\pi}$ in MT-MFG. Then we have:

$$
\begin{aligned}
&\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}-\boldsymbol{\mu}^{\boldsymbol{\pi}}_h\|_1\\
=&\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}-\Gamma^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\|_1+\|\Gamma^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}(p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})-\boldsymbol{\mu}^{\boldsymbol{\pi}}_h\|_1\\
\leq&\varepsilon+\sum_{\widetilde{w}\in[W]}\sum_{s_h^{\widetilde{w}}}|\sum_{s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}}}p^{\widetilde{w},\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}(s_{h-1}^{\widetilde{w}})\pi^{\widetilde{w}}_{h-1}(a_{h-1}^{\widetilde{w}}|s_{h-1}^{\widetilde{w}})\mathbb{P}^{\widetilde{w}}_{h-1}(s_h^{\widetilde{w}}|s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}},p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\\
&\qquad-\sum_{s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}}}\mu^{\widetilde{w},\boldsymbol{\pi}}_{h-1}(s_{h-1}^{\widetilde{w}})\pi^{\widetilde{w}}_{h-1}(a_{h-1}^{\widetilde{w}}|s_{h-1}^{\widetilde{w}})\mathbb{P}^{\widetilde{w}}_{h-1}(s_h^{\widetilde{w}}|s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}},\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h-1})|\\
\leq&\varepsilon+\sum_{\widetilde{w}\in[W]}\sum_{s_h^{\widetilde{w}}}\sum_{s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}}}|p^{\widetilde{w},\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}(s_{h-1}^{\widetilde{w}})-\mu^{\widetilde{w},\boldsymbol{\pi}}_{h-1}(s_{h-1}^{\widetilde{w}})|\pi^{\widetilde{w}}_{h-1}(a_{h-1}^{\widetilde{w}}|s_{h-1}^{\widetilde{w}})\mathbb{P}^{\widetilde{w}}_{h-1}(s_h^{\widetilde{w}}|s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}},p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})\\
&+\sum_{\widetilde{w}\in[W]}\sum_{s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}}}\mu^{\widetilde{w},\boldsymbol{\pi}}_{h-1}(s_{h-1}^{\widetilde{w}})\pi^{\widetilde{w}}_{h-1}(a_{h-1}^{\widetilde{w}}|s_{h-1}^{\widetilde{w}})\sum_{s_h^{\widetilde{w}}}|\mathbb{P}^{\widetilde{w}}_{h-1}(s_h^{\widetilde{w}}|s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}},p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}})-\mathbb{P}^{\widetilde{w}}_{h-1}(s_h^{\widetilde{w}}|s_{h-1}^{\widetilde{w}},a_{h-1}^{\widetilde{w}},\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h-1})|\\
\leq&\varepsilon+(1+\boldsymbol{L}_T)\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_{h-1}}-\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h-1}\|_1\\
\leq&\frac{(1+\boldsymbol{L}_T)^h-1}{\boldsymbol{L}_T}\varepsilon.
\end{aligned}
$$

Upper Bound of Approximation Error Recall the definition of value functions in Def. H.1 and Def. H.2. We focus on the $(w,1)$-agent which takes a potentially deviated policy $\widetilde{\pi}^w$ while the others do not, and we are interested in provide an upper bound for the value difference $J^{(w,1)}_{\bar{\boldsymbol{M}}}(\widetilde{\boldsymbol{\nu}})-J^{(w,1)}_{\boldsymbol{M}}(\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w;\boldsymbol{\pi})$, which will be useful to characterize the sub-optimality of lifted policy $\boldsymbol{\nu}$.

We start from step $h = H$, following the choice of N^w in Eq. (23),

$$\begin{aligned}
&\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[V^{(w,1),\widetilde{\boldsymbol{\nu}}}_{\bar{\boldsymbol{M}},H}(s^{(w,1)}_H;\boldsymbol{s}_H) - V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},H}(s^{(w,1)}_H;\boldsymbol{\mu}^{\boldsymbol{\pi}}_H)]\\
=&\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[r^w_H(s^{(w,1)}_H,a^{(w,1)}_H;p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_H}) - r^w_H(s^{(w,1)}_H,a^{(w,1)}_H;\boldsymbol{\mu}^{\boldsymbol{\pi}}_H)]\\
\le&\boldsymbol{L}_r\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_H} - \boldsymbol{\mu}^{\boldsymbol{\pi}}_H\|_1] \le \boldsymbol{L}_r(2\delta + \frac{(1+\boldsymbol{L}_T)^H - 1}{\boldsymbol{L}_T}\varepsilon) =: \varepsilon_H.
\end{aligned}$$

For $h < H$, we have:

$$\begin{aligned}
&\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[V^{(w,1),\widetilde{\boldsymbol{\nu}}}_{\bar{\boldsymbol{M}},h}(s^{(w,1)}_h;\boldsymbol{s}_h) - V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h}(s^{(w,1)}_h;\boldsymbol{\mu}^{\boldsymbol{\pi}}_h)]\\
=&\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[r^w_h(s^{(w,1)}_h,a^{(w,1)}_h;p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}) - r^w_h(s^{(w,1)}_h,a^{(w,1)}_h;\boldsymbol{\mu}^{\boldsymbol{\pi}}_h) + \sum_{\boldsymbol{s}_{h+1}}\mathbb{P}_{\bar{\boldsymbol{M}}}(\boldsymbol{s}_{h+1}|\boldsymbol{s}_h,\boldsymbol{a}_h)V^{(w,1),\widetilde{\boldsymbol{\nu}}}_{\bar{\boldsymbol{M}},h+1}(s^{(w,1)}_{h+1};\boldsymbol{s}_{h+1})\\
&- \sum_{s^{(w,1)}_{h+1}}\mathbb{P}^w_h(s^{(w,1)}_{h+1}|s^{(w,1)}_h,a^{(w,1)}_h,\boldsymbol{\mu}^{\boldsymbol{\pi}}_h)V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h+1}(s^{(w,1)}_{h+1};\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h+1})]\\
\le&\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\boldsymbol{L}_r\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h} - \boldsymbol{\mu}^{\boldsymbol{\pi}}_h\|_1 + \sum_{\boldsymbol{s}_{h+1}}\mathbb{P}_{\bar{\boldsymbol{M}}}(\boldsymbol{s}_{h+1}|\boldsymbol{s}_h,\boldsymbol{a}_h)\Big(V^{(w,1),\widetilde{\boldsymbol{\nu}}}_{\bar{\boldsymbol{M}},h+1}(s^{(w,1)}_{h+1};\boldsymbol{s}_{h+1}) - V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h+1}(s^{(w,1)}_{h+1};\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h+1})\Big)\\
&+ \sum_{\boldsymbol{s}_{h+1}}\mathbb{P}_{\bar{\boldsymbol{M}}}(\boldsymbol{s}_{h+1}|\boldsymbol{s}_h,\boldsymbol{a}_h)V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h+1}(s^{(w,1)}_{h+1};\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h+1})\\
&- \sum_{s^{(w,1)}_{h+1}}\mathbb{P}^w_h(s^{(w,1)}_{h+1}|s^{(w,1)}_h,a^{(w,1)}_h,\boldsymbol{\mu}^{\boldsymbol{\pi}}_h)V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h+1}(s^{(w,1)}_{h+1};\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h+1})]\\
\le&\varepsilon_{h+1} + \mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\boldsymbol{L}_r\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h} - \boldsymbol{\mu}^{\boldsymbol{\pi}}_h\|_1 + \sum_{\boldsymbol{s}_{h+1}}\mathbb{P}_{\bar{\boldsymbol{M}}}(\boldsymbol{s}_{h+1}|\boldsymbol{s}_h,\boldsymbol{a}_h)V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h+1}(s^{(w,1)}_{h+1};\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h+1})\\
&- \sum_{s^{(w,1)}_{h+1}}\mathbb{P}^w_h(s^{(w,1)}_{h+1}|s^{(w,1)}_h,a^{(w,1)}_h,\boldsymbol{\mu}^{\boldsymbol{\pi}}_h)V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h+1}(s^{(w,1)}_{h+1};\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h+1})] \qquad \text{(By induction from } h+1)\\
=&\varepsilon_{h+1} + \mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\boldsymbol{L}_r\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h} - \boldsymbol{\mu}^{\boldsymbol{\pi}}_h\|_1 +\\
&+ \sum_{s^{(w,1)}_{h+1}}\Big(\mathbb{P}^w_h(s^{(w,1)}_{h+1}|s^{(w,1)}_h,s^{(w,1)}_h,p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h}) - \mathbb{P}^w_h(s^{(w,1)}_{h+1}|s^{(w,1)}_h,a^{(w,1)}_h,\boldsymbol{\mu}^{\boldsymbol{\pi}}_h)\Big)V^{(w,1),\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w}_{\boldsymbol{M},h+1}(s^{(w,1)}_{h+1};\boldsymbol{\mu}^{\boldsymbol{\pi}}_{h+1})]\\
\le&\varepsilon_{h+1} + (\boldsymbol{L}_r+\boldsymbol{L}_T)\mathbb{E}_{\bar{\boldsymbol{M}},\widetilde{\boldsymbol{\nu}}}[\|p^{\widetilde{\boldsymbol{\nu}}}_{\boldsymbol{s}_h} - \boldsymbol{\mu}^{\boldsymbol{\pi}}_h\|_1] \le \varepsilon_{h+1} + (\boldsymbol{L}_r+\boldsymbol{L}_T)(2\delta + \frac{(1+\boldsymbol{L}_T)^h - 1}{\boldsymbol{L}_T}\varepsilon) =: \varepsilon_h.
\end{aligned}$$

where we use $\mathbb{P}_{\bar{\boldsymbol{M}}}(\boldsymbol{s}_{h+1}|\boldsymbol{s}_h,\boldsymbol{a}_h) := \prod_{w\in[W]}\prod_{n\in[N]}\mathbb{P}^w_h(s^{(w,1)}_{h+1}|s^{(w,1)}_h,a^{(w,1)}_h,p_{\boldsymbol{s}_h})$ to denote the dynamics in MF-SAG. Therefore, for $h = 1$, note that $\bar{\boldsymbol{M}}$ and $\boldsymbol{M}$ have the same initial distribution, and we have:

$$J^{(w,1)}_{\bar{\boldsymbol{M}}}(\widetilde{\boldsymbol{\nu}}) - J^{(w,1)}_{\boldsymbol{M}}(\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w;\boldsymbol{\pi}) \le 2\delta(\boldsymbol{L}_r+\boldsymbol{L}_T)H + 2(\boldsymbol{L}_r+\boldsymbol{L}_T)H\frac{(1+\boldsymbol{L}_T)^H - 1}{\boldsymbol{L}_T}\varepsilon.$$

Given an $\widetilde{\varepsilon}$ NE policy in $\boldsymbol{M}$, denoted by $\boldsymbol{\pi}$, consider the lifted policy $\boldsymbol{\nu}$ and a deviation policy $\widetilde{\boldsymbol{\nu}}$ agrees with $\boldsymbol{\nu}$ except that it takes some $\widetilde{\pi}^w$ for agent with index $(w,1)$. By choosing $\delta = \frac{\varepsilon_0}{8(\boldsymbol{L}_r+\boldsymbol{L}_T)H}$ and $\varepsilon = \varepsilon_0/\Big(4(\boldsymbol{L}_r+\boldsymbol{L}_T)H\frac{(1+\boldsymbol{L}_T)^H-1}{\boldsymbol{L}_T}\Big)$, we have:

$$\begin{aligned}
\max_{\widetilde{\pi}^w} J^{(w,1)}_{\bar{\boldsymbol{M}}}(\widetilde{\boldsymbol{\nu}}) - J^{(w,1)}_{\bar{\boldsymbol{M}}}(\boldsymbol{\nu}) \le& \max_{\widetilde{\pi}^w} J^{(w,1)}_{\bar{\boldsymbol{M}}}(\widetilde{\boldsymbol{\nu}}) - J^{(w,1)}_{\bar{\boldsymbol{M}}}(\boldsymbol{\nu}) - \Big(\max_{\widetilde{\pi}^w} J^{(w,1)}_{\boldsymbol{M}}(\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w;\boldsymbol{\pi}) - J^{(w,1)}_{\boldsymbol{M}}(\boldsymbol{\pi};\boldsymbol{\pi})\Big) + \widetilde{\varepsilon}\\
\le& 2\max_{\widetilde{\pi}^w}|J^{(w,1)}_{\bar{\boldsymbol{M}}}(\widetilde{\boldsymbol{\nu}}) - J^{(w,1)}_{\boldsymbol{M}}(\boldsymbol{\pi}^{-w}\circ\widetilde{\pi}^w;\boldsymbol{\pi})| + \widetilde{\varepsilon} \le \varepsilon_0 + \widetilde{\varepsilon}.
\end{aligned}$$

To satisfy the requirements in δ and ε_0, we need:

$$\forall w\in[W],\quad N^w \ge O((\boldsymbol{L}_r+\boldsymbol{L}_T)^2W^2H^3\Big(\frac{(1+\boldsymbol{L}_T)^H-1}{\boldsymbol{L}_T}\Big)^2\frac{\sum_{w\in[W]}S^w}{\varepsilon_0^2}\log\frac{2(\boldsymbol{L}_r+\boldsymbol{L}_T)(1+\boldsymbol{L}_T)HWS_{\max}}{\varepsilon_0}).$$

□

I. Basic Lemma

I.1. Lemma from (Huang et al., 2023)

Lemma I.1 (Lem. D.4 in (Huang et al., 2023))**.** *Let $X_1, X_2, ...$ be a sequence of random variable taking value in $[0, C]$ for some $C \geq 1$. Define $\mathcal{F}_k = \sigma(X_1, .., X_{k-1})$ and $Y_k = \mathbb{E}[X_k|\mathcal{F}_k]$ for $k \geq 1$. For any $\delta > 0$, we have:*

$$\Pr(\exists n \sum_{k=1}^n X_k \leq 3\sum_{k=1}^n Y_k + C\log\frac{1}{\delta}) \leq \delta, \quad \Pr(\exists n \sum_{k=1}^n Y_k \leq 3\sum_{k=1}^n X_k + C\log\frac{1}{\delta}) \leq \delta.$$

Lemma I.2 (Lem. 4.6 in (Huang et al., 2023))**.** *Under Assump. B, given two arbitrary model M and $\widetilde{M}$, and two policies π and $\widetilde{\pi}$, we have:*

$$\begin{aligned}|\Delta_M(\widetilde{\pi}, \pi) - \Delta_{\widetilde{M}}(\widetilde{\pi}, \pi)| \leq & \mathbb{E}_{\widetilde{\pi}, M(\pi)}[\sum_{h=1}^H \|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu^\pi_{\widetilde{M},h})\|_1] \\ & + (2L_r H + 1)\mathbb{E}_{\pi, M(\pi)}[\sum_{h=1}^H \|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu^\pi_{\widetilde{M},h})\|_1]. \end{aligned} \tag{24}$$

I.2. Other Lemma

Lemma I.3 (Density Difference Lemma)**.** *Given arbitrary Multi-Type Mean-Field MDPs $\boldsymbol{M}$ and $\boldsymbol{M}'$, and two arbitrary policies $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$, for any $h \in [H]$, we have:*

$$\begin{aligned}\|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M}',h}\|_1 \leq & \|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M},h-1}\|_1 + W \cdot d_{\infty,1}(\boldsymbol{\pi}, \boldsymbol{\pi}') \\ & + \sum_{w \in [W]} \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{M}(\boldsymbol{\pi})}[\|\mathbb{P}^w_{\boldsymbol{M},h}(\cdot|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}) - \mathbb{P}^w_{\boldsymbol{M}',h}(\cdot|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M}',h-1})\|_1].\end{aligned}$$

Proof. For any $w \in [W]$, we have:

$$\begin{aligned}& \|\mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h} - \mu^{w,\boldsymbol{\pi}'}_{\boldsymbol{M}',h}\|_1 \\ = & |\sum_{s^w_h} \Big(\sum_{s^w_{h-1}, a^w_{h-1}} \mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h}(s^w_{h-1})\pi^w_{h-1}(a^w_{h-1}|s^w_{h-1})\mathbb{P}^w_{\boldsymbol{M},h}(s^w_h|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}) \\ & - \sum_{s^w_{h-1}, a^w_{h-1}} \mu^{w,\boldsymbol{\pi}'}_{\boldsymbol{M}',h}(s^w_h)\pi'^w_{h-1}(a^w_{h-1}|s^w_{h-1})\mathbb{P}^w_{\boldsymbol{M}',h}(s^w_h|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M}',h-1})\Big)| \\ \leq & |\Big(\sum_{s^w_{h-1}} \mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h}(s^w_{h-1}) - \mu^{w,\boldsymbol{\pi}'}_{\boldsymbol{M}',h}(s^w_{h-1})\Big) \sum_{a^w_{h-1}} \pi'^w_{h-1}(a^w_{h-1}|s^w_{h-1}) \sum_{s_h} \mathbb{P}^w_{\boldsymbol{M}',h}(s^w_h|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M}',h-1})| \\ & + |\sum_{s^w_{h-1}} \mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h}(s^w_{h-1}) \sum_{a^w_{h-1}} \Big(\pi^w_{h-1}(a^w_{h-1}|s^w_{h-1}) - \pi'^w_{h-1}(a^w_{h-1}|s^w_{h-1})\Big) \sum_{s^w_h} \mathbb{P}^w_{\boldsymbol{M}',h}(s^w_h|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M}',h-1})| \\ & + \sum_{s^w_{h-1}, a^w_{h-1}} \mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h}(s^w_h)\pi^w_{h-1}(a^w_{h-1}|s^w_{h-1}) \sum_{s^w_h} |\mathbb{P}^w_{\boldsymbol{M},h}(s^w_h|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}) - \mathbb{P}^w_{\boldsymbol{M}',h}(s^w_h|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M}',h-1})| \\ & \hfill \text{(Assump. B)} \\ \leq & \|\mu^{w,\boldsymbol{\pi}}_{\boldsymbol{M},h-1} - \mu^{w,\boldsymbol{\pi}'}_{\boldsymbol{M}',h-1}\|_1 + d_{\infty,1}(\boldsymbol{\pi}, \boldsymbol{\pi}') + \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{M}(\boldsymbol{\pi})}[\|\mathbb{P}^w_{\boldsymbol{M},h}(\cdot|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}) - \mathbb{P}^w_{\boldsymbol{M}',h}(\cdot|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M}',h-1})\|_1]\end{aligned}$$

By repeating the above discussion for every $w \in [W]$, we have:

$$\begin{aligned}& \|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M}',h}\|_1 \\ \leq & \|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1} - \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M},h-1}\|_1 + W \cdot d_{\infty,1}(\boldsymbol{\pi}, \boldsymbol{\pi}') \\ & + \sum_{w \in [W]} \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{M}(\boldsymbol{\pi})}[\|\mathbb{P}^w_{\boldsymbol{M},h}(\cdot|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h-1}) - \mathbb{P}^w_{\boldsymbol{M}',h}(\cdot|s^w_{h-1}, a^w_{h-1}, \boldsymbol{\mu}^{\boldsymbol{\pi}'}_{\boldsymbol{M}',h-1})\|_1].\end{aligned}$$

□

Lemma I.4. *Given two model $\boldsymbol{M}$ and $\boldsymbol{M}'$ and a policy $\boldsymbol{\pi}$, for any $h \in [H], w \in [W]$, we have:*

$$\|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h+1} - \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M}',h+1}\|_1 \leq \sum_{w\in[W]} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{M}(\boldsymbol{\pi})}[\sum_{h'=1}^{h} \|\mathbb{P}^w_{\boldsymbol{M},h'}(\cdot|\cdot,\cdot,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h'}) - \mathbb{P}^w_{\boldsymbol{M}',h'}(\cdot|\cdot,\cdot,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M}',h'})\|_1]. \tag{25}$$

Besides, under Assump. B, we have:

$$\|\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h+1} - \boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M}',h+1}\|_1 \leq \sum_{w\in[W]} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{M}(\boldsymbol{\pi})}[\sum_{h'=1}^{h} (1+\boldsymbol{L}_T)^{h-h'} \|\mathbb{P}^w_{\boldsymbol{M},h'}(\cdot|\cdot,\cdot,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h'}) - \mathbb{P}^w_{\boldsymbol{M}',h'}(\cdot|\cdot,\cdot,\boldsymbol{\mu}^{\boldsymbol{\pi}}_{\boldsymbol{M},h'})\|_1]. \tag{26}$$

Proof. By applying Lem. I.3 to the case when $\boldsymbol{\pi} = \boldsymbol{\pi}'$, and combining with Assump. D, we finish the proof. □

Lemma I.5. *Given two model M and M', and two arbitrary policies π and π', for any $h \in [H]$, we have:*

$$\begin{aligned}&\|\mu^{\pi}_{M,h} - \mu^{\pi'}_{M',h}\|_1\\ \leq&\|\mu^{\pi}_{M,h-1} - \mu^{\pi'}_{M,h-1}\|_1 + d_{\infty,1}(\pi,\pi') + \mathbb{E}_{\pi,M(\pi)}[\|\mathbb{P}_{M,h}(\cdot|s_{h-1},a_{h-1},\mu^{\pi}_{M,h-1}) - \mathbb{P}_{M',h}(\cdot|s_{h-1},a_{h-1},\mu^{\pi'}_{M',h-1})\|_1].\end{aligned}$$

Moreover, as a special case when $\pi = \pi'$, we have:

$$\|\mu^{\pi}_{M,h+1} - \mu^{\pi}_{M',h+1}\|_1 \leq \mathbb{E}_{\pi,M(\pi)}[\sum_{h'=1}^{h} \|\mathbb{P}_{M,h'}(\cdot|s_{h'},a_{h'},\mu^{\pi}_{M,h'}) - \mathbb{P}_{M',h'}(\cdot|s_{h'},a_{h'},\mu^{\pi}_{M',h'})\|_1]. \tag{27}$$

Besides, under Assump. B, we have:

$$\|\mu^{\pi}_{M,h+1} - \mu^{\pi}_{M',h+1}\|_1 \leq \mathbb{E}_{\pi,M(\pi)}[\sum_{h'=1}^{h} (1+L_T)^{h-h'} \|\mathbb{P}_{M,h'}(\cdot|s_{h'},a_{h'},\mu^{\pi}_{M,h'}) - \mathbb{P}_{M',h'}(\cdot|s_{h'},a_{h'},\mu^{\pi}_{M,h'})\|_1]. \tag{28}$$

Proof. The proof is simply completed by setting $W = 1$ in Lem. I.4. □

Lemma I.6 (Concentration w.r.t. l_1-distance)**.** *Given a discrete domain $\mathcal{X}$ and a distribution p on $\mathcal{X}$, suppose we draw N i.i.d. samples $\{x^n\}_{n\in[N]}$ from p and provide an estimation $\widehat{p} \in \Delta(\mathcal{X})$ with $\widehat{p}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x^n = x)$, then for any $\delta \in (0,1)$ and $\varepsilon > 0$, as long as $N \geq \max\{\frac{2|\mathcal{X}|}{\varepsilon^2}, \frac{2}{\varepsilon^2}\log\frac{2}{\delta}\}$, we have:*

$$\mathbb{P}(\|p - \widehat{p}\|_1 \geq \varepsilon) \leq \delta.$$

Proof. We first provide an upper bound for $\mathbb{E}[\|p - \widehat{p}\|_1]$:

$$\mathbb{E}[\|p - \widehat{p}\|_1] = \sum_{x\in\mathcal{X}} \mathbb{E}[|p(x) - \widehat{p}(x)|] \leq \sum_{x\in\mathcal{X}} \sqrt{\mathbb{E}[|p(x) - \widehat{p}(x)|^2]} = \sqrt{\frac{1}{N}} \sum_{x\in\mathcal{X}} \sqrt{p(x)(1-p(x))} \leq \frac{1}{2}\sqrt{\frac{|\mathcal{X}|}{N}}.$$

where we use the fact that $\widehat{p}(x)$ is a Bernoulli random variable with mean $p(x)$ and variance $\frac{1}{N}p(x)(1-p(x))$.

Next, and note that deviation of any x^n will only result in $2/N$ deviation of $\|p - \widehat{p}\|_1$. By McDiarmid's inequality, for any ε, we have:

$$\mathbb{P}(\|p - \widehat{p}\|_1 \geq \frac{\varepsilon}{2} + \frac{1}{2}\sqrt{\frac{2|\mathcal{X}|}{N}}) \leq \mathbb{P}(|\|p - \widehat{p}\|_1 - \mathbb{E}[\|p - \widehat{p}\|_1]| \geq \frac{\varepsilon}{2}) \leq 2\exp(-\frac{N}{2}\varepsilon^2),$$

By assigning appropriate values for N, we finish the proof. □

Lemma I.7. *[Model Difference Lemma] For any policies $\widetilde{\pi}$, π and π', and any bounded functions $f_1, f_2, ..., f_H \in \{f | f : \mathcal{S} \times \mathcal{A} \to [0,1]\}$,*

(i) Given any two MF-MDPs M and M', we have:

$$\left|\mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} f_h(s_h,a_h)] - \mathbb{E}_{\widetilde{\pi},M'(\pi')}[\sum_{h=1}^{H} f_h(s_h,a_h)]\right| \leq H \cdot \mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_M(\cdot|s_h,a_h,\mu^{\pi}_{M,h}) - \mathbb{P}_{M'}(\cdot|s_h,a_h,\mu^{\pi'}_{M',h})\|_1].$$

(ii) Given any two PAMs $\ddot{M}$ and $\ddot{M}'$, we have:

$$\left|\mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\sum_{h=1}^{H} f_h(s_h,a_h)] - \mathbb{E}_{\widetilde{\pi},\ddot{M}'(\pi')}[\sum_{h=1}^{H} f_h(s_h,a_h)]\right| \leq H \cdot \mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}}(\cdot|s_h,a_h,\pi) - \ddot{\mathbb{P}}_{\ddot{M}'}(\cdot|s_h,a_h,\pi')\|_1].$$

Proof. We first proof (ii). We use $\mu^{\pi}_{\ddot{M}(\pi'),h}$ to denote the density induced by π in model $\ddot{M}$ given π' as the reference policy.

$$\begin{aligned}
&\|\mu^{\widetilde{\pi}}_{\ddot{M}(\pi),h} - \mu^{\widetilde{\pi}}_{\ddot{M}'(\pi'),h}\|_1 \\
=&|\sum_{s_h}\Big(\sum_{s_{h-1},a_{h-1}} \mu^{\widetilde{\pi}}_{\ddot{M}(\pi),h-1}(s_{h-1})\widetilde{\pi}(a_{h-1}|s_{h-1})\ddot{\mathbb{P}}_{\ddot{M}}(s_h|s_{h-1},a_{h-1},\pi) \\
&\quad - \sum_{s_{h-1},a_{h-1}} \mu^{\widetilde{\pi}}_{\ddot{M}'(\pi'),h-1}(s_{h-1})\widetilde{\pi}(a_{h-1}|s_{h-1})\ddot{\mathbb{P}}_{\ddot{M}'}(s_h|s_{h-1},a_{h-1},\pi')\Big)| \\
\leq&|\sum_{s_h}\sum_{s_{h-1},a_{h-1}} \mu^{\widetilde{\pi}}_{\ddot{M}(\pi),h-1}\widetilde{\pi}(a_{h-1}|s_{h-1})(\ddot{\mathbb{P}}_{\ddot{M}}(s_h|s_{h-1},a_{h-1},\pi) - \ddot{\mathbb{P}}_{\ddot{M}'}(s_h|s_{h-1},a_{h-1},\pi'))| \\
&+|\sum_{s_h}\sum_{s_{h-1},a_{h-1}} (\mu^{\widetilde{\pi}}_{\ddot{M}(\pi),h-1}(s_{h-1}) - \mu^{\widetilde{\pi}}_{\ddot{M}'(\pi'),h-1}(s_{h-1}))\widetilde{\pi}(a_{h-1}|s_{h-1})\ddot{\mathbb{P}}_{\ddot{M}'}(s_h|s_{h-1},a_{h-1},\pi')| \\
\leq&\mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\|\ddot{\mathbb{P}}_{\ddot{M}}(\cdot|s_{h-1},a_{h-1},\pi) - \ddot{\mathbb{P}}_{\ddot{M}'}(\cdot|s_{h-1},a_{h-1},\pi')\|_1] + \|\mu^{\widetilde{\pi}}_{\ddot{M}(\pi),h-1} - \mu^{\widetilde{\pi}}_{\ddot{M}'(\pi'),h-1}\|_1 \\
\leq&\mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\sum_{h'=1}^{h-1} \|\ddot{\mathbb{P}}_{\ddot{M}}(\cdot|s_{h'},a_{h'},\pi) - \ddot{\mathbb{P}}_{\ddot{M}'}(\cdot|s_{h'},a_{h'},\pi')\|_1].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left|\mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\sum_{h=1}^{H} f_h(s_h,a_h)] - \mathbb{E}_{\widetilde{\pi},\ddot{M}'(\pi')}[\sum_{h=1}^{H} f_h(s_h,a_h)]\right| &\leq \sum_{h=1}^{H} \|\mu^{\widetilde{\pi}}_{\ddot{M}(\pi),h} - \mu^{\widetilde{\pi}}_{\ddot{M}'(\pi'),h}\|_1 \\
&\leq H \cdot \mathbb{E}_{\widetilde{\pi},\ddot{M}(\pi)}[\sum_{h=1}^{H} \|\ddot{\mathbb{P}}_{\ddot{M}}(\cdot|s_h,a_h,\pi) - \ddot{\mathbb{P}}_{\ddot{M}'}(\cdot|s_h,a_h,\pi')\|_1].
\end{aligned}$$

The proof for (i) can be directly obtained by replacing $\ddot{\mathbb{P}}_{\ddot{M},h}(\cdot|\cdot,\cdot,\pi)$ and $\ddot{\mathbb{P}}_{\ddot{M}',h}(\cdot|\cdot,\cdot,\pi')$ with $\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h})$ and $\mathbb{P}_{M',h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M',h})$.

□

Lemma I.8. *Given three arbitrary models $M, \widetilde{M}, \bar{M}$, and two arbitrary policies $\pi, \widetilde{\pi}$, we have:*

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\pi},\widetilde{M}(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{\widetilde{M},h}(\cdot|s_h,a_h,\mu^{\pi}_{\widetilde{M},h}) - \mathbb{P}_{\bar{M},h}(\cdot|s_h,a_h,\mu^{\pi}_{\bar{M},h})\|_1] \\
\leq&\mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|s_h,a_h,\mu^{\pi}_{M,h}) - \mathbb{P}_{\bar{M},h}(\cdot|s_h,a_h,\mu^{\pi}_{\bar{M},h})\|_1] \\
&+ (H+1) \cdot \mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|s_h,a_h,\mu^{\pi}_{M,h}) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h,a_h,\mu^{\pi}_{\widetilde{M},h})\|_1].
\end{aligned} \tag{29}$$

Proof. By applying Lem. I.7 with $f_h(s_h, a_h) = \|\mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h}) - \mathbb{P}_{\bar{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\bar{M},h})\|_1$, we have:

$$
\begin{aligned}
&\mathbb{E}_{\widetilde{\pi},\widetilde{M}(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h}) - \mathbb{P}_{\bar{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\bar{M},h})\|_1] \\
\leq &\mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h}) - \mathbb{P}_{\bar{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\bar{M},h})\|_1] \\
&+ H \cdot \mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^{\pi}_{M,h}) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h})\|_1] \\
\leq &\mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^{\pi}_{M,h}) - \mathbb{P}_{\bar{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\bar{M},h})\|_1] \\
&+ (H+1)\mathbb{E}_{\widetilde{\pi},M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu^{\pi}_{M,h}) - \mathbb{P}_{\widetilde{M},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h})\|_1]. \qquad (30)
\end{aligned}
$$

□

J. Details of Experiments

J.1. Algorithm Design

In the following, we provide the missing algorithm details for Sec. 6.

Algorithm 7: A Heuristic Oracle-Efficient NE Finding

Input: Model Class $\mathcal{M}$; Accuracy level $\varepsilon_0, \widetilde{\varepsilon}, \bar{\varepsilon}$; Confidence level δ; Batch size T
$\mathcal{M}^1 \leftarrow \mathcal{M}$; $\delta_0 \leftarrow \frac{\delta}{\log_2 |\mathcal{M}|+1}$
$\forall M \in \mathcal{M}^1$, compute (one of) its NE policy $\pi^{\text{NE}}_M \leftarrow \texttt{NE\_Compute}(M)$.
for $k = 1, 2, ...,$ **do**
  if $\exists M^k \in \mathcal{M}^k$, *s.t.* $\max_{\widetilde{M} \in \mathcal{M}^k} |\mathcal{B}^{\varepsilon_0}_{\pi^{NE}_{M^k}}(\widetilde{M}, \mathcal{M}^k)| \leq \frac{|\mathcal{M}^k|}{2}$ **then**
    $\mathcal{M}^{k+1} \leftarrow \texttt{ModelElim\_Exp}(\pi^{\text{NE}}_{M^k}, \mathcal{M}^k, \widetilde{\varepsilon}, \delta_0, T)$.
  end
  else
    $M^k \leftarrow \arg\max_{M \in \mathcal{M}^k} |\mathcal{B}^{\varepsilon_0}_{\pi^{\text{NE}}_M}(M, \mathcal{M}^k)|$,
    $\pi^{\text{NE}}_{\text{Br}} \leftarrow \pi^{\text{NE}}_{M^k}$,
    $\mathcal{M}^{k+1} \leftarrow \texttt{ModelElim\_Exp}(\pi^{\text{NE}}_{\text{Br}}, \mathcal{M}^k, \widetilde{\varepsilon}, \delta_0, T)$.
    if $M^k \in \mathcal{M}^{k+1}$ **then return** $\pi^{\text{NE}}_{M^k}$;
  end
end

Here `ModelElim_Exp` is the same algorithm as Alg. 2 except that we replac Line 4 with:

$$\widetilde{\pi}^t \leftarrow \arg\max_{\widetilde{\pi} \in \Pi^{\text{NE}}} \max_{M, M' \in \bar{\mathcal{M}}^t} \mathbb{E}_{\widetilde{\pi}, M(\pi)}[\sum_{h=1}^{H} \|\mathbb{P}_{M,h}(\cdot|\cdot, \cdot, \mu^{\pi}_{M,h}) - \mathbb{P}_{M',h}(\cdot|\cdot, \cdot, \mu^{\pi}_{M',h})\|_1],$$

In another word, we only consider policies from $\Pi^{\text{NE}} := \{\pi^{\text{NE}}_M\}_{M \in \bar{\mathcal{M}}}$, including the NE policies of models in $\bar{\mathcal{M}}$.

J.2. Experiment Setup

Environments We consider the linear style MFG, such that

$$\mathbb{P}(s_{h+1}|s_h, a_h, \mu_h) = \frac{|\phi(s_h, a_h)^\top G(\mu_h)\psi(s_{h+1})|}{\sum_{s_{h+1}} |\phi(s_h, a_h)^\top G(\mu_h)\psi(s_{h+1})|},$$

where $\phi \in \mathbb{R}^{d_\phi}$ and $G(\cdot) \in \mathbb{R}^{d_\phi \times d_\psi}$ are known but $\psi \in \mathbb{R}^{d_\psi}$ are unknown. Note that our environment is different from linear model in Prop. D.12, where features are self-normalized. We choose $H = 3, S = 100, A = 50$ and $d_\phi = d_\psi = 5$, where the number of states and actions is much larger than the feature dimension. We consider a model set with $|\mathcal{M}| = 200$.

To construct the environment, for each h, we first generate a random matrix $\Phi_h \in \mathbb{R}^{SA \times d_\phi}$ using as feature $\phi(s_h, a_h)$, and generate another random matrix $U_h \in \mathbb{R}^{S \times d_\phi d_\psi}$, and define the function $G_h(\mu_h)$ by

$$\forall \mu_h \in \Delta(\mathcal{S}_h), \quad G_h(\mu_h) := (\mu_h^\top U_h).\text{reshape}(d_\phi, d_\psi) \in \mathbb{R}^{d_\phi, d_\psi}.$$

After that, we generate 200 random matrices $\{\Psi^i_h\}_{i \in [200]}$ with $\Psi^i_h \in \mathbb{R}^{d_\psi \times S}$ as the next feature function. Then, the model class is specified by $\mathcal{M} := \{(\Phi_h, U_h, \Psi^i_h)\}$. In order to make the model elimination process more challenging, $\{\Psi^i_h\}_{i=2,...,200}$ is generated by randomly perturbing from Ψ^1_h, i.e.:

$$\Psi^i_h = (1 - \beta)\widetilde{\Psi}^i_h + \beta \Psi^1_h,$$

where $\widetilde{\Psi}^i_h$ is a random matrix independent w.r.t. Ψ^1_h and $\beta \sim \text{Uniform}(0, 0.1)$. In this way, the difference between models in $\mathcal{M}$ will be small and harder to distinguish.

Training Procedure We construct 5 model classes $\mathcal{M}^1, ..., \mathcal{M}^5$ with different Φ, U to increase the randomness in experiments. For each model class $\mathcal{M}^i$, we repeat 5 trials, where in each trial, we first randomly select one model from $\mathcal{M}^i$ as the true model, and run Alg. 7 for model elimination.

We set ε=1e-3, i.e. we want to find a 1e-3-approximate NE. Besides, we set batch size $T = 50$, $\delta = 0.001$. For the `NE_Oracle` in Alg. 7, we implement it by repeatedly update

$$\pi_{i+1} \leftarrow (1-\alpha)\pi_i + \alpha \text{BestReponse}(\pi_i; M). \tag{31}$$

where $\alpha = 0.02$, and BestReponse$(\pi_i; M)$ return the policy maximizing the NE gap of π_i in M. We stop the update process as long as $\mathcal{E}_M^{\text{NE}}(\pi_i) \leq$ 5e-4.

Experiments Results We provide our experiment results in Fig. 2. On the LHS, we report the number of uneliminated models verses the number of trajectories consumed, and as we can see, our algorithm can eliminate unqualified models very quickly. The total consumed trajectories is much less than the number of states actions $SA = 100 * 50 = 5000$.

On the RHS[7], we report the normalized worst case NE Gap w.r.t. the remaining models. At each iteration t, we compute the NE gap for every uneliminated model's NE policy, and pick out the largest one denoted as Gap_t. The normalized gap is defined to be $\frac{\text{Gap}_t}{\text{Gap}_0}$, where the normalization term Gap_0 is the maximal NE gap at the beginning of the algorithm, i.e. the worst NE gap without starting the algorithm. As we can see, our algorithm can gradually eliminate inaccurate models and return the (approximate) NE.

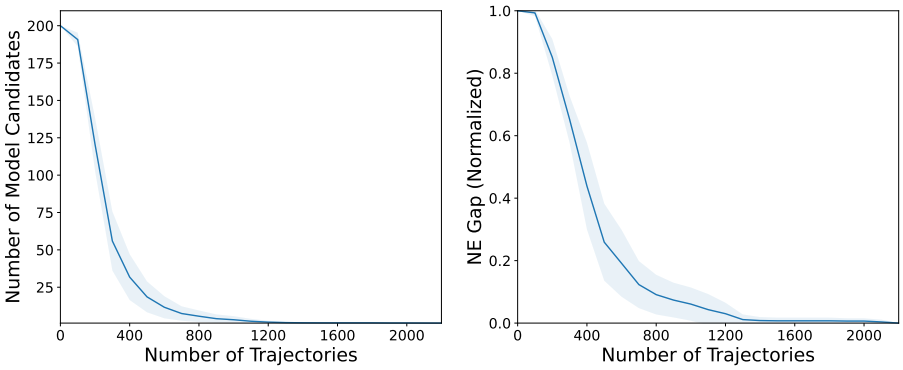

Figure 2. Experiment results in linear style MFG. We report the number of remaining models and the normalized maximal NE Gap by the NE policies of remaining models during the model elimination process. Error bars correspond to 95% confidence intervals.

[7]In the RHS sub-plot of Fig. 2, we set the normalized NE Gap to 0 as long as it is lower than 1e-3