On the Statistical Efficiency of Mean Field RL with General Function Approximation

Jiawei Huang **Batuhan Yardim** **Niao He**
Department of Computer Science
ETH Zurich
{jiawei.huang, alibatuhan.yardim, niao.he}@inf.ethz.ch

Abstract

In this paper, we study the statistical efficiency of Reinforcement Learning in Mean-Field Control (MFC) and Mean-Field Game (MFG) with general function approximation. We introduce a new concept called Mean-Field Model-Based Eluder Dimension (MBED), which subsumes a rich family of Mean-Field RL problems. Additionally, we propose algorithms based on Optimistic Maximal Likelihood Estimation, which can return an ε-optimal policy for MFC or an ε-Nash Equilibrium policy for MFG, with sample complexity polynomial w.r.t. relevant parameters and independent of the number of states, actions and the number of agents. Notably, our results only require a mild assumption of Lipschitz continuity on transition dynamics comparing with previous works. Finally, in the tabular setting, given the access to a generative model, we establish an exponential lower bound for MFC setting, while providing a novel sample-efficient model elimination algorithm to approximate equilibrium in MFG setting. Our results reveal a fundamental separation between RL for single-agent, MFC, and MFG from the sample efficiency perspective.

1 Introduction

Multi-Agent Reinforcement Learning (MARL) is a fundamental model that addresses how multiple autonomous agents cooperate or compete with each other in a shared environment, and it is widely applied for practical problems in many areas, including autonomous driving [53], finance [37], and robotics control [29]. Although MARL has attracted growing attention in nowadays RL research [26, 33, 34], when the number of agents is in hundreds or thousands, MARL already becomes challenging. However, in scenarios where agents exhibit high symmetry, like humans in crowds or individual cars in the traffic flow, the Mean-Field theory can be employed to approximate the system dynamics, which results in the Mean-Field RL (MFRL) setting. In MFRL, the interactions within large populations are modeled by the additional dependence of state density of agents (population distribution) of the transition function in the Mean-Field Markov Decision Process (MF-MDP). Such mathematical model has achieved success in various domains, including economics [15, 4], finance [11], industrial engineering [18], etc.

Depending on the objectives, MFRL can be divided into two categories: Mean-Field Control (MFC) and Mean-Field Game (MFG) [36, 28, 10]. MFC, similar to the single-agent RL, aims to find a policy maximizing the expected return, while MFG focuses on identifying the Nash Equilibrium (NE) policy, where no agent has the incentive to deviate. Compared with the single-agent RL, one of the main challenges in MFRL is the exploration in the joint space of state, action, and state density, especially given that the density belongs to an infinite and continuous space. Due to this challenge, existing literature primarily focuses on the tabular setting, and most results rely on strong assumptions like contractivity [24, 59], monotonicity [50, 49, 20], or population-independent dynamics [41, 22].

16th European Workshop on Reinforcement Learning (EWRL 2023).

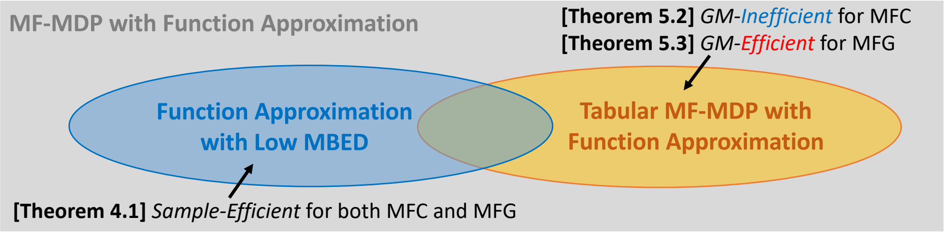
MF-MDP with Function Approximation
[Theorem 5.2] GM-Inefficient for MFC
[Theorem 5.3] GM-Efficient for MFG
Function Approximation with Low MBED
Tabular MF-MDP with Function Approximation
[Theorem 4.1] Sample-Efficient for both MFC and MFG

Figure 1: **Highlight of Main Results (Dependence on Lipschitz factors are omitted)**: "*Sample-Efficient*" means the problem can be solved with polynominal samples by Data Collection Process in Def. 2.2. "*GM-Efficient/Inefficient*" means the problem can/cannot be solved with polynomial queries to the Generative Model (GM) in Def. 5.1.

Recent work [48] considered function approximation in MFC, while they did not study MFG setting (see Sec. 1.1 for more comparison). Overall, efficiently solving MFRL, especially MFG, with general function approximation under mild structural assumptions remains an open problem.

In this paper, our goal is to investigate the statistical efficiency of Mean Field RL with model-based function approximation. We summarize our main results in Fig. 1 and highlight our contributions in threefold. Firstly, in Sec. 3, we introduce a new complexity measure called Model-Based Eluder Dimension for MFRL (MBED), and contribute concrete classes of MFRL problems with low MBED, including (generalized) linear MF-MDP, deterministic transition with Gaussian noise, and etc. To our knowledge, the understanding of MBED is limited in both single-agent RL and MARL literature.

Secondly, in Sec. 4, we develop efficient model-learning algorithms for MFRL based on Optimism-Maximal Likelihood Estimation (O-MLE), and prove that, if the model class has bounded MBED, polynomial sample complexity can be achieved in both MFC and MFG setting. As we will see, the dependence on state density in transition function causes unique challenges, and we overcome them by establishing close connections between model class complexity, MLE error, and the learning error for MFC and MFG objectives. Notably, our results do not require strong structural assumptions like previous work [24, 50, 49, 20]. As a by-product, our results imply sample efficiency for model-based RL with function approximation in the single-agent setting, which might be of independent interest.

Given the results in Sec. 4, where we identify a structural condition under which both MFC and MFG can be solved efficiently, one natural question arises regarding whether the two objectives share the same statistical efficiency or inefficiency in general. As our third contribution, in Sec. 5, we provide evidence for the exponential separation between MFC and MFG given the generative model (GM) [24, 20]. In tabular setting with function approximation, for MFC, even with access to GM, we establish an exponential lower bound for finding an near-optimal solution. In contrast, for MFG, by identifying the special property implied by "local alignment" (Lem. 5.4), we propose a novel model elimination algorithm, which can approximate the NE policy with polynomial queries to GM. To the best of our knowledge, this is the first result indicating the separation between MFC and MFG from the sample efficiency perspective. In general, finding NE or establishing sample complexity lower bound for NE is believed to be hard in many MARL/MFRL cases, we believe our results provide important insights on these directions.

1.1 Closely Related Work

For the lack of space, we only highlight the most related works here and defer the others to Appx. A.2. In general, the theoretical understanding of MFRL in the finite horizon setting is still limited, especially in terms of statistical efficiency. We present and compare with several lines of work.

Finite-horizon MFG. The finite-horizon framework considered here is closely related to Lasry-Lions games [50, 49, 22], where continuous-time dynamics were analyzed without exploration considerations under monotonicity assumptions on rewards. While [22] proves discrete time convergence for rewards admitting a potential, they have not considered exploration. Our work focuses on understanding the fundamental exploration guarantees and bottlenecks associated with finite-horizon MFC and MFG, hence applies also to MFG and MFC not satisfying restrictive conditions. Working in a similar setting, [20] requires a planning oracle that can return a trajectory for arbitrary density, even if it can not be induced by any policy.

Comparison to stationary MFG equilibrium. Alternative to the finite-horizon formulation, there exists work on the stationary MFG formulation where one aims to find policies which keep the population stationary [3, 24, 59, 63, 16]. In this formulation, results typically require strong Lipschitz continuity assumptions as well as non-vanishing regularization [24, 3, 16]. In [24], strong Lipschitz assumptions (like Assump. 1 and 2) are required, which is unclear when it can be true. Furthermore, they also assume a stronger generator model which can infer the trajectory for arbitrary density function. Besides, their regret bound will have a dependence on the covering time, which potentially scales with $O(|\mathcal{S}||\mathcal{A}|)$, and their convergence rate scales at $\mathcal{O}(\varepsilon^{-5})$, while ours can achieve the optimal rate at $\mathcal{O}(\varepsilon^{-2})$. They also require assumptions on the gap of value functions induced by the epsilon net of the density. [59] has the same smoothness requirements, which might not be avoidable [16, 63]. Furthermore, the common formalization that the transition and reward function in each step conditions on the stationary density instead of the density evolved across time could be a limitation.

Comparing with statistical efficiency results for MFC. In terms of statistical efficiency considerations, a similar work in MFC has been [48]. But our results capture a different learnable function class, with some overlap. As our advantages, our low MBED can capture multimodal transition distribution (e.g. by linear setting), while their algorithm and analysis is specified for deterministic transition with random noise (unimodal transition distribution). Besides, our framework can include some speical cases in [48]. For those near-deterministic transition functions modeled by Gaussian Process with additive uncorrelated Gaussian noise, ours can predict its sample efficiency, as a result of our Prop. 3.5 and the equivalence between Eluder Dimension and Information Gain in RKHS space [32]. As our insufficiency, [48] can handle the cases when the noise is sub-gaussian besides pure gaussian, as long as they can get access to the full information of the noise, while it is unclear whether such function class has low MBED. It would be an interesting direction to propose more general complexity measure and algorithms, which can unify the frameworks in both papers together.

Besides, we also analyze the MFG setting while they only focused on MFC. Furthermore, they only considers with the deterministic policy class, while in many cases, the optimal policy or the equilibrium policy can only be stochastic. In contrast, we allow stochastic policy, do not assume the knowledge of density (this is also reflected in the policy we compete with).

Other MFG/MFC settings. There also exists a variety of different settings in which MFG formalism has been utilized, for instance in linear quadratic MFG [25] and MFG on graphs [62, 23]. [5] studies a unified view of MFG and MFC, however, they do not take the evolution of density into consideration and do not provide guarantees for the non-tabular setting. Several works on MFC also work on the lifted MDP where population state is observable [12]. In our work, we do not assume the observability of the population.

2 Preliminary

2.1 Setting and Frequently Used Notations

We consider the finite-horizon Mean-Field Markov Decision Process (MF-MDP) specified by a tuple $M := (\mu_1, \mathcal{S}, \mathcal{A}, H, \mathbb{P}_T, \mathbb{P}_r)$. Here μ_1 is the fixed initial distribution known to the learner, $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively, which can be discrete or continuous and compact. Besides, we assume the state action space are the same for each step h, i.e., $\mathcal{S}_h = \mathcal{S}$ and $\mathcal{A}_h = \mathcal{A}$ for all h. $\mathbb{P}_T := \{\mathbb{P}_{T,h}\}_{h=1}^H$ and $\mathbb{P}_r := \{\mathbb{P}_{r,h}\}_{h=1}^H$ are the transition and (normalized) deterministic reward function, with $\mathbb{P}_{T,h} : \mathcal{S}_h \times \mathcal{A}_h \times \Delta(S_h) \rightarrow \Delta(S_{h+1})$ and $\mathbb{P}_{r,h} : \mathcal{S}_h \times \mathcal{A}_h \times \Delta(S_h) \rightarrow [0, \frac{1}{H}]$. To be concise in analysis, we assume that the reward function is known, but our techniques can be extended when it is unknown. We use M^* to denote the true model with transition function $\mathbb{P}_{T^*}$.

In this paper, we only consider the non-stationary Markov policy $\pi := \{\pi_1, ..., \pi_H\}$ with $\pi_h : \mathcal{S}_h \rightarrow \Delta(\mathcal{A}_h)$, $\forall\ h \in [H]$. Starting from the initial state $s_1 \sim \mu_1$ until the fixed final state s_{H+1} is reached, the trajectory is generated by:

$$\forall h \in [H] \quad a_h \sim \pi_h(\cdot|s_h),\ s_{h+1} \sim \mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu^\pi_h),\ r_h \sim \mathbb{P}_{r,h}(\cdot|s_h, a_h, \mu^\pi_{M,h}),\ \mu^\pi_{M,h+1} = \Gamma^\pi_{M,h}(\mu^\pi_{M,h}),$$

$$\text{with}\quad \Gamma^\pi_{M,h}(\mu_h)(\cdot) := \sum\!\!\!\!\!\!\!\int_{s_h,a_h} \mu_h(s_h)\pi(a_h|s_h)\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu_h), \tag{1}$$

where we use $\mu^\pi_{M,h}$ to denote the density induced by π in M and $\Gamma^\pi_{M,h} : \Delta(\mathcal{S}_h) \rightarrow \Delta(\mathcal{S}_{h+1})$ is an mapping from densities in step h to step $h+1$ under M and π. We will use bold font $\boldsymbol{\mu} := \{\mu_1, ..., \mu_H\}$ to denote the collection of density for all time steps, and use $r_h(s_h, a_h, \mu_h)$ to denote the expected reward at s_h, a_h, μ_h. Besides, we denote $V^\pi_{M,h}(\cdot; \boldsymbol{\mu})$ to be the value function at step h if the agent deploys policy π in model M conditioning on $\boldsymbol{\mu}$, defined by:

$$V^\pi_{M,h}(s_h; \boldsymbol{\mu}) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, \mu_{h'})\Big|_{a_{\widetilde{h}}\sim\pi_{\widetilde{h}},\ s_{\widetilde{h}+1}\sim\mathbb{P}_{T,\widetilde{h}}(\cdot|s_{\widetilde{h}},a_{\widetilde{h}},\mu_{\widetilde{h}}),\ \ \forall \widetilde{h} \geq h}\right].$$

We use $J_M(\pi; \boldsymbol{\mu}) := \mathbb{E}_{s_1\sim\mu_1}[V^\pi_{M,1}(s_1; \boldsymbol{\mu})]$ to denote the expected return of policy π in model M conditioning on $\boldsymbol{\mu}$. When the policy is specified, we use $\boldsymbol{\mu}^\pi_M := \{\mu^\pi_{M,1}, ..., \mu^\pi_{M,H}\}$ to denote the collection of mean fields w.r.t. π. We will omit $\boldsymbol{\mu}$ and use $J_M(\pi)$ in shorthand when $\boldsymbol{\mu} = \boldsymbol{\mu}^\pi_M$. For the simplicity, in the rest of the paper, we use $\mathbb{E}_{\pi,M|\boldsymbol{\mu}}[\cdot] := \mathbb{E}\left[\cdot\Big|_{\substack{s_1\sim\mu_1 \\ \forall h\geq 1,\quad a_h\sim\pi_h(\cdot|s_h) \\ s_{h+1}\sim\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu_h)}}\right]$ as a shortnote of the expectation over trajectories induced by π under transition $\mathbb{P}_{T,h}(\cdot|\cdot,\cdot,\mu_h)$, and we omit the conditional density $\boldsymbol{\mu}$ if $\boldsymbol{\mu} = \boldsymbol{\mu}^\pi_M$. As examples, $V^\pi_{M,h}(s_h; \boldsymbol{\mu}) = \mathbb{E}_{\pi,M|\boldsymbol{\mu}}[\sum_{h'=h}^H r(s_{h'}, a_{h'}, \mu_{h'})|s_h]$ and $J_M(\pi) = \mathbb{E}_{\pi,M}[\sum_{h=1}^H r(s_{h'}, a_{h'}, \mu^\pi_{M,h'})]$.

Given a measure space $(\Omega, \mathcal{F})$ and two probability measures P and Q defined on $(\Omega, \mathcal{F})$, we denote $\mathbb{TV}(P,Q)$ (or $\|P - Q\|_{\mathbb{TV}}$)$:= \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$ as the total variation distance, and denote $\mathbb{H}(P,Q) := \sqrt{1 - \sum\!\!\!\!\!\!\int_x \sqrt{P(x)Q(x)}}$ as the Hellinger distance. In general, we have $\sqrt{2}\mathbb{H}(P,Q) \geq \mathbb{TV}(P,Q)$. When Ω is countable, $\mathbb{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1$, where $\|\cdot\|_1$ is the l_1-distance.

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

$$\mathcal{E}_{\text{Opt}}(\widehat{\pi}^*_{\text{Opt}}) \leq \varepsilon. \tag{2}$$

Mean-Field Game: In MFG, we instead want to find a NE policy s.t., when all the agents follow that same policy, no agent tends to deviate it for better policy value. We denote $\Delta_M(\widetilde{\pi}, \pi) := J_M(\widetilde{\pi}; \boldsymbol{\mu}^\pi_M) - J_M(\pi; \boldsymbol{\mu}^\pi_M)$ given a model M, and denote $\mathcal{E}_{\text{NE}}(\pi) := \max_{\widetilde{\pi}} \Delta_{M^*}(\widetilde{\pi}, \pi)$, which is also known as the exploitability. The NE in M^* is defined to be the policy π^*_{NE} satisfying $\mathcal{E}_{\text{NE}}(\pi^*_{\text{NE}}) = 0$. Our MFG objective is to find an approximate NE $\widehat{\pi}^*_{\text{NE}}$ such that:

$$\mathcal{E}_{\text{NE}}(\widehat{\pi}^*_{\text{NE}}) \leq \varepsilon. \tag{3}$$

2.2 Assumptions

In this paper, we consider the general function approximation setting, where the learner can get access to a model class $\mathcal{M}$ satisfying the following assumptions.

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

Assumption B (Lipschitz Continuity)**.** For arbitrary $h \in [H], s_h \in \mathcal{S}, a_h \in \mathcal{A}$ and arbitrary valid density $\mu_h, \mu'_h \in \Delta(\mathcal{S})$, and arbitrary model $M := (\mathbb{P}_T, \mathbb{P}_r) \in \mathcal{M}$, we have: [1]

$$\mathbb{H}(\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu_h), \mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu'_h)) \leq L_T \cdot \|\mu_h - \mu'_h\|_{\mathbb{TV}}. \tag{4}$$

$$\|\mathbb{P}_{r,h}(\cdot|s_h,a_h,\mu_h) - \mathbb{P}_{r,h}(\cdot|s_h,a_h,\mu'_h)\|_{\mathbb{TV}} \leq L_r \cdot \|\mu_h - \mu'_h\|_{\mathbb{TV}} \tag{5}$$

Assumption C (Existence of NE)**.** For any $M \in \mathcal{M}$, there exists at least one NE policy.

Although we treat Assump. C as an assumption, in Prop. 2.1 below, we show it is implied by Assump. B for discrete environments, and the proof can be generalized to many continuous cases.

[1] Here we consider the Lipschitz continuity w.r.t. $\mathbb{H}$ just in order to coordinate with our formulation of MBED in Def. 3.3. In fact, Eq. (4) can be relaxed to $\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu_h), \mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu'_h)\|_{\mathbb{TV}} \leq L_T \cdot \|\mu_h - \mu'_h\|_{\mathbb{TV}}$ if we only consider $\mathbb{TV}$ distance in Eq. (6).

Proposition 2.1 (Existence of NE in MFG; Informal Version of Prop. E.7)**.** *For every MF-MDP with discrete $\mathcal{S}$ and $\mathcal{A}$, satisfying Assump. B, there exists at least one NE policy.*

We also note that the existence of NE is established in previous literature [52] under the same conditions as our Prop. E.7. Our contribution here is a different proof based on the conjugate function and non-expansiveness of the proximal point operator. Moreover, [52] studied infinite-horizon MDP with discounted reward, which is different from our setting.

Besides, we formalize the Data Collection Process in the following.

Definition 2.2 (Data Collection Process (DCP))**.** We assume the environment consists of an extremely large number of agents and a central controller (our algorithm/learner), and there is a probe agent `Agt`, whose observation we can receive. The central controller can compute an arbitrary policy tuple $(\widetilde{\pi}, \pi)$, where π and $\widetilde{\pi}$ are not necessarily the same, distribute $\widetilde{\pi}$ to `Agt` but π to the others, and receive the trajectory of `Agt` following $\widetilde{\pi}$ under $\mathbb{P}_{T^*,h}(\cdot|\cdot,\cdot,\mu_h^\pi)$ and $\mathbb{P}_{r,h}(\cdot|\cdot,\cdot,\mu_h^\pi)$.

Our formulation above is reasonable, since the deviation of one agent only causes neglectable perturbation on density. Besides, it is much weaker than the assumption in [24, 20], which requires a planning oracle that can return a trajectory conditioning on arbitrary (even unachievable) density.

3 Model-Based Eluder Dimension for Mean-Field RL

We first introduce our definition for Model-Based Eluder-Dimension in MFRL.

Definition 3.1 (α-weakly-ε-independent sequence)**.** Denote $\mathcal{X} := \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$ to be the joint space of state, action and state density. Let $\mathbf{D} : \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \to [0, C]$ be a distribution distance measure bounded by some constant C. Given a function class $\mathcal{F} \subset \{f : \mathcal{X} \to \Delta(\mathcal{S})\}$, a fixed $\alpha \geq 1$ and a sequence of data points $x_1, x_2, ..., x_n \in \mathcal{X}$, we say x is α-weakly-ε-independent of $x_1, ..., x_n$ w.r.t. $\mathcal{F}$ and $\mathbf{D}$ if there exists $f_1, f_2 \in \mathcal{F}$ such that $\sum_{i=1}^n \mathbf{D}^2(f_1, f_2)(x_i) \leq \varepsilon^2$ but $\mathbf{D}(f_1, f_2)(x) > \alpha\varepsilon$.

Definition 3.2 (Longest α-weakly-ε-independent sequence)**.** We use $\mathrm{dimE}_\alpha(\mathcal{F}, \mathbf{D}, \varepsilon)$ to denote the the longest sequence $\{x_i\}_{i=1}^n \in \mathcal{X}$, such that for some $\varepsilon' \geq \varepsilon$, x_i is α-weakly-ε'-independent of $\{x_1, ..., x_{i-1}\}$ for all $i \in [n]$ w.r.t. $\mathcal{F}$ and $\mathbf{D}$.

Definition 3.3 (Model-Based Eluder-Dimension in MFRL)**.** Given a model class $\mathcal{M}$, $\alpha \geq 1$ and $\varepsilon > 0$, the Model-Based Eluder Dimension in MFRL (abbr. MBED) of $\mathcal{M}$ is defined to be:

$$\mathrm{dimE}_\alpha(\mathcal{M}, \varepsilon) := \max_{h \in [H]} \min_{\mathbf{D} \in \{\mathbb{TV}, \mathbb{H}\}} \mathrm{dimE}_\alpha(\mathcal{M}_h, \mathbf{D}, \varepsilon). \tag{6}$$

We only consider $\mathbf{D}$ to be $\mathbb{TV}(P, Q)$ or $\mathbb{H}(P, Q)$, mainly because of our MLE-based loss function. With slightly abuse of notation, the $\mathcal{M}$ (or $\mathcal{M}_h$) here refers the collection of transition functions of models in $\mathcal{M}$. The main difference comparing with value function approximation setting [51, 32] is that, because the output of model functions are distributions instead of scalar, we use distance measure to compute the model prediction difference. Besides, we use $\alpha\varepsilon$ as threshold instead of ε, which does not lead to a fundamentally different complexity measure, but simplifies the process to absorb some practical examples into our framework. Also note that $\mathrm{dimE}_{\alpha_1}(\mathcal{F}, \varepsilon) \leq \mathrm{dimE}_{\alpha_2}(\mathcal{F}, \varepsilon)$ for $\alpha_1 \geq \alpha_2$, because any α_1-weakly-ε-independent sequence must be α_2-weakly-ε-independent.

Comparison with previous work regarding MBED To our knowledge, only few literature has focused on Model-Based Eluder Dimension (MBED). [46] requires additional assumption that, given two transition distributions in the function class, the difference between their induced future value function is Lipschitz continuous w.r.t. the their mean difference, which is quite restrictive. In a more recent work, [38] presented extension of MBED to general bounded metrics, however, their results still depend on the number of states actions, and concrete examples with low MBED are not provided.

Concrete Examples Next, we introduce some concrete examples with low MBED, and defer formal statements and their proofs to Appx. B.2. The first one is generalized from the linear MDP in single-agent RL [35]. In Appx. B.2, we also include a linear mixture type model, and other more general examples, such as, kernel MF-MDP and the generalized linear MF-MDP. However, since the output of the model function is a probability distribution rather than a scalar, low TV-distance between predictions does not necessarily imply they are uniformly close for each output dimension, which causes technical challenges. To overcome it, we utilize data-dependent sign functions to pave ways to establish the connection between the prediction error and the elliptical potential lemma.

Proposition 3.4 (Low-Rank MF-MDP with Known Representation; Informal Version of Prop. B.4)**.** *Given a feature $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \rightarrow \mathbb{R}^d$ and a function class Ψ, the model class $\mathcal{P}_\Psi := \{\mathbb{P}_\psi | \mathbb{P}_\psi(s'|s,a,\mu) := \phi(s,a,\mu)^\top \psi(s'),\ \psi \in \Psi\}$ has $dimE_\alpha(\mathcal{P}_\Psi, \mathbb{TV}, \varepsilon) = \widetilde{O}(d)$ for $\alpha \geq 1$.*

The second example is deterministic transition with random noise, in order to accommodate the function class in [48] (see a detailed comparison in Sec. 1.1). Here we consider the Hellinger distance because given two Gaussian distribution $P \sim \mathcal{N}(\mu_P, \Sigma)$ and $Q \sim \mathcal{N}(\mu_Q, \Sigma)$ with the same covariance, $\mathbb{H}(P,Q) = 1 - \exp(-\frac{1}{8}\|\mu_P - \mu_Q\|^2_{\Sigma^{-1}})$. Therefore, with the connection between $\mathbb{H}(P,Q)$ and the l_2-distance between their mean value, we are able to subsume more important model classes into low MBED framework, as we state below.

Proposition 3.5. *[Deterministic Transition with Gaussian Noise] Suppose $\mathcal{S} \subset \mathbb{R}^d$. Given a function class $\mathcal{G} \subset \{g | g : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \times \mathbb{N}^* \rightarrow \mathbb{R}\}$ and convert it to $\mathcal{F}_\mathcal{G} := \{f_g | f_g(\cdot,\cdot,\cdot) := [g(\cdot,\cdot,\cdot,1), ..., g(\cdot,\cdot,\cdot,d)]^\top \in \mathbb{R}^d,\ g \in \mathcal{G}\}$. Consider the model class $\mathcal{P}_\mathcal{G} := \{\mathbb{P}_f | \mathbb{P}_f(\cdot|s,a,\mu) \sim f(s,a,\mu) + \mathcal{N}(0,\Sigma), f \in \mathcal{F}_\mathcal{G}\}$, where $\mathcal{N}(0,\Sigma)$ is the Gaussian noise with $\Sigma := \text{Diag}(\sigma, ..., \sigma)$. For $\varepsilon \leq 0.3$, we have $dimE_{\sqrt{2}}(\mathcal{P}_\mathcal{G}, \mathbb{H}, \varepsilon) \leq \overline{dimE}(\mathcal{F}_\mathcal{G}, 4\sigma\varepsilon)$, $dimE_{\sqrt{2d}}(\mathcal{P}_\mathcal{G}, \mathbb{H}, \varepsilon) \leq \overline{dimE}(\mathcal{G}, 4\sigma\varepsilon)$, where $\overline{dimE}$ is the Eluder Dimension for scalar or vector-valued functions [51, 46].*

4 Learning in Mean-Field RL: An Optimistic-MLE Approach

In this section, we introduce our O-MLE based algorithm for MFRL. We highlight the algorithm design and main results in Sec. 4.1, and introduce key techniques to results in Sec. 4.2.

4.1 Main Algorithm and Results

We provide our main algorithm in Alg. 1, where we omit the rewards in samples to avoid redundancy in analysis. Similar to previous work for function approximation setting in single-agent RL [32, 19] or MARL [26], we mainly focus on the statistical complexity and leave the computational efficiency to future work. Our algorithm is based on Optimistic Maximial Likelihood Estimation. The algorithm includes two parts: policy selection (Line 8-14) and data collection (Line 4-7). In each iteration k, we fit the model with data $\mathcal{Z}^1, ..., \mathcal{Z}^k$ collected so far and construct a model confidence set $\widehat{\mathcal{M}}^k$. The confidence level is carefully chosen, so that with high probability, we can ensure $M^* \in \widehat{\mathcal{M}}^k$ for all k.

In MFC, similar to the single-agent setting, we pick π^{k+1} to be the policy achieving the maximal total return among models in the confidence set, and then use it to collect new samples for exploration. In the end, we use `Regret2PAC` conversion algorithm (Alg. 3, deferred to Appx. D.3) to select policy.

For MFG, the learning process is slightly more complicated. For the policy selection part, we compute two policies. We first randomly pick M^{k+1} from $\widehat{\mathcal{M}}^k$, and compute its equilibrium policy π^{k+1} to be our guess for the equilibrium of the true model M^*. Next, we find a model $\widetilde{M}^{k+1}$ and an adversarial policy $\widetilde{\pi}^{k+1}$, which result in an optimistic estimation for $\mathcal{E}_{\text{NE}}(\pi^{k+1})$. Besides, for the data collection part, in addition to the trajectories generated by deploying π^{k+1}, we also collect trajectories sampled by policy $\widetilde{\pi}^{k+1}$ conditioning on the density induced by π^{k+1}. As we will explain in Lem. 4.6, those additional samples are necessary to control the estimation error of exploitability. Finally, we return the policy with the minimal optimistic exploitability among $\{\pi^{k+1}\}_{k=1}^K$.

We state our main results below, and defer its formal version (Thm. D.5 and Thm. D.6) and the proofs to Appx. D. As a side contribution, our results can recover sample complexity in single-agent RL by letting $L_T, L_r \rightarrow 0$, which is only studied by few literatures.

Theorem 4.1 (Main Results (Informal))**.** *Under Assump.A, B and C, by choosing[2] $K = \widetilde{O}\Big((1 + L_r H)^2(1 + L_T H)^2\Big(\frac{(1+L_T)^H - 1}{L_T}\Big)^2 \frac{dimE_\alpha(\mathcal{M}, \varepsilon_0)}{\varepsilon^2}\Big)$, with $\varepsilon_0 = O(\frac{L_T \varepsilon}{\alpha H(1+L_r H)(1+L_T H)((1+L_T)^H - 1)})$,*

(i) for the MFC branch, after consuming HK trajectories in Alg. 1 and additional $O(\frac{1}{\varepsilon^2}\log^2\frac{1}{\delta})$ trajectories in Alg. 3, w.p. $1 - 5\delta$, we have $\mathcal{E}_{Opt}(\widehat{\pi}^*_{Opt}) \leq \varepsilon$.

(ii) for the MFG branch, after consuming $2HK$ trajectories, w.p. $1 - 3\delta$, we have $\mathcal{E}_{NE}(\widehat{\pi}^*_{NE}) \leq \varepsilon$.

[2] We omit log-dependence on ε, δ, dimE, $|\mathcal{M}|$, H and Lipschitz factors in $\widetilde{O}$.

Exponential Dependence on L_T: As we can see, there is an exponential dependence of L_T in sample complexity, while similar result has been reported in previous literatures [48]. Besides a trivial observation that the exponential factor reduces to constant when $L_T = O(\frac{1}{H})$, in Appx. D.1, we introduce Assump. D about the contraction of Γ^π operator, which is frequently considered in previous literature [3, 63]. In the full version theorem, we show that with that additional assumption, the sample complexity only depends on a contractive factor without exponential terms.

Algorithm 1: A General O-MLE Learning Framework for Mean-Field RL

```
1  Input: Model function class $\mathcal{M}$; $\varepsilon, \delta, K$.
2  Initialize: Randomly pick $\pi^1$ and $\tilde{\pi}^1$; $\mathcal{Z}^k \leftarrow \{\}, \forall k \in [K]$.
3  for $k = 1, 2, ..., K$ do
4      for $h = 1, ..., H$ do
5          Sample $z_h^k := \{s_h^k, a_h^k, s'^k_{h+1}\}$ with $(\pi^k, \pi^k)$; $\mathcal{Z}^k \leftarrow \mathcal{Z}^k \cup z_h^k$.
6          if MFG then Sample $\tilde{z}_h^k := \{\tilde{s}_h^k, \tilde{a}_h^k, \tilde{s}'^k_{h+1}\}$ with $(\tilde{\pi}^k, \pi^k)$; $\mathcal{Z}^k \leftarrow \mathcal{Z}^k \cup \tilde{z}_h^k$ ;
7      end
8      For each $M \in \mathcal{M}$, define:
       $l^k_{\text{MLE}}(M) := \sum_{i=1}^k \sum_{h=1}^H \log \mathbb{P}_{T,h}(s'^i_{h+1}|s^i_h, a^i_h, \mu^{\pi^i}_{M,h}) + \underbrace{\log \mathbb{P}_{T,h}(\tilde{s}'^i_{h+1}|\tilde{s}^i_h, \tilde{a}^i_h, \mu^{\pi^i}_{M,h})}_{\text{MFG only}}$.
9      $\widehat{\mathcal{M}}^k \leftarrow \{M \in \mathcal{M} | l^k_{\text{MLE}}(M) \geq \max_{M \in \mathcal{M}} l^k_{\text{MLE}}(M) - \log \frac{2|\mathcal{M}|KH}{\delta}\}$.
10     if MFC then $\pi^{k+1}, M^{k+1} \leftarrow \arg\max_{\pi, M \in \widehat{\mathcal{M}}^k} J_M(\pi; \boldsymbol{\mu}^\pi_M)$ ;
11     if MFG then
12         Randomly pick $M^{k+1}$ from $\widehat{\mathcal{M}}^k$; Find a NE of $M^{k+1}$ denoted as $\pi^{k+1}$.
13         $\tilde{\pi}^{k+1}, \widetilde{M}^{k+1} \leftarrow \arg\max_{\tilde{\pi}; M \in \widehat{\mathcal{M}}^k} \Delta_M(\tilde{\pi}, \pi^{k+1})$.
14     end
15 end
16 if MFC then return $\hat{\pi}^*_{\text{Opt}} \leftarrow \text{Regret2PAC}(\{\pi^{k+1}\}_{k=1}^K, \varepsilon, \delta)$ ;
17 if MFG then return $\hat{\pi}^*_{\text{NE}} \leftarrow \pi^{k^*_{\text{NE}}}$ with $k^*_{\text{NE}} \leftarrow \min_{k \in [K]} \Delta_{\widetilde{M}^{k+1}}(\tilde{\pi}^{k+1}, \pi^{k+1})$ ;
```

4.2 Proof Sketch

The high-level ideas for proving Thm. 4.1 can be mainly divided into two parts. Firstly, we provide an upper bound for the accumulative model prediction error by the model-based eluder dimension, which we further connect with our learning objective in the second step.

Step 1: Upper Bound Model Prediction Error with MBED First of all, in Thm. 4.2 below, we show that, with high probability, models in $\widehat{\mathcal{M}}^k$ predict well under the distribution of data collected so far. We defer the proof to Appx. C.

Theorem 4.2. *[Guarantees for MLE] By running Alg. 1 with any $\delta \in (0, 1)$, with probability $1 - \delta$, for all $k \in [K]$, we have $M^* \in \widehat{\mathcal{M}}^k$; for each $M \in \widehat{\mathcal{M}}^k$ with transition $\mathbb{P}_T$ and any $h \in [H]$:*

$$\sum_{i=1}^k \mathbb{E}_{\pi^i, M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h, a^i_h, \mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h, a^i_h, \mu^{\pi^i}_{M^*,h}))] \leq 2\log(\frac{2|\mathcal{M}|KH}{\delta}).$$

Besides, for MFG branch, we additionally have:

$$\sum_{i=1}^k \mathbb{E}_{\tilde{\pi}^i, M^*|\boldsymbol{\mu}^{\pi^i}_{M^*}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\tilde{s}^i_h, \tilde{a}^i_h, \mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|\tilde{s}^i_h, \tilde{a}^i_h, \mu^{\pi^i}_{M^*,h}))] \leq 2\log(\frac{2|\mathcal{M}|KH}{\delta}).$$

The key difficulty in Mean-Field setting is the dependence of density in transition function. Since we do not know $\mu^\pi_{M^*,h}$, in Line 8 in Alg. 1, we compute the likelihood conditioning on $\mu^\pi_{M,h}$, which is accessible for each M. Therefore, in Thm. 4.2, we can only guarantee M aligns with M^* conditioning on their own density $\boldsymbol{\mu}^\pi_M$ and $\boldsymbol{\mu}^\pi_{M^*}$, respectively. However, to ensure low MBED can indeed capture important practical models, the MBED in Def. 3.3 is established on shared density, which is also the main reason we additional consider Hellinger distance in Assump. B. To close this gap, in Thm. 4.3 below, we present how the model difference conditioning on the same or different densities can be converted to each other. The proof is defered to Appx. D.

Theorem 4.3 (Model Difference Conversion; Short Version of Thm. D.3)**.** *Given two model M and $\widetilde{M}$ with transition $\mathbb{P}_T$ and $\mathbb{P}_{\widetilde{T}}$, respectively, and an arbitrary policy π, under Assump. B, we have:*

$$\begin{aligned}
&\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]\\
&\qquad\qquad\leq(1+L_TH)\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}], \qquad (7)\\
&\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]\\
&\qquad\qquad\leq\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}(1+L_T)^{H-h}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]. \qquad (8)
\end{aligned}$$

In the final lemma, we show if the model predicts well in history, then the growth rate of the accumulative error on new data can be controlled by MBED. We defer the proof to Appx. B.3.

Lemma 4.4. *Under the condition as Def. 3.1, consider a fixed $f^*\in\mathcal{F}$, and suppose we have a sequence $\{f_k\}_{k=1}^K\in\mathcal{F}$ and $\{x_k\}_{k=1}^K\subset\mathcal{S}\times\mathcal{A}\times\Delta(\mathcal{S})$, s.t., for all $k\in[K]$, $\sum_{i=1}^{k-1}\boldsymbol{D}^2(f_k,f^*)(x_i)\leq\beta$, then for any $\varepsilon>0$, we have $\sum_{k=1}^K\boldsymbol{D}(f_k,f^*)(x_k)=O(\sqrt{\beta K dimE_\alpha(\mathcal{M},\varepsilon)}+\alpha K\varepsilon)$.*

Step 2: Relating Learning Objectives with Model Prediction Error First of all, we provide the simulation lemma for Mean-Field Control setting.

Lemma 4.5. *[Simulation Lemma for MFC] Given an arbitrary model M with transition function $\mathbb{P}_T$, and an arbitrary policy π, under Assump. B, we have:*

$$|J_{M^*}(\pi)-J_M(\pi)|\leq\mathbb{E}_{\pi,M^*}[\sum_{h=1}^{H}(1+L_rH)\|\mathbb{P}_{T^*,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h})-\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}].$$

By Thm. 4.2 and Eq. (7) in Thm. 4.3, with high probability, all the models in $\widehat{\mathcal{M}}_k$ will agrees with each other on the dataset D^k conditioning on the same density $\mu^{\pi^1}_{M^*},...,\mu^{\pi^k}_{M^*}$. On good concentration events, the condition for Lem. 4.4 is satisfied, and as a result of Thm. 4.5 and Eq. (8), we can upper bound the accumulative sub-optimal gap $\sum_{k=1}^K\mathcal{E}_{\text{Opt}}(\pi^{k+1})$. With the regret to PAC convertion process in Alg. 3, we can establish the sample complexity guarantee in Thm. 4.1.

For MFG, we first provide an upper bound for $\mathcal{E}_{\text{NE}}(\pi^{k+1})$. On the event of $M^*\in\widehat{\mathcal{M}}^{k+1}$, we have:

$$\begin{aligned}
\mathcal{E}_{\text{NE}}(\pi^{k+1})=&\max_\pi\Delta_{M^*}(\pi,\pi^{k+1})\leq\Delta_{\widetilde{M}^{k+1}}(\widetilde{\pi}^{k+1},\pi^{k+1})\leq\Delta_{\widetilde{M}^{k+1}}(\widetilde{\pi}^{k+1},\pi^{k+1})-\Delta_{M^{k+1}}(\widetilde{\pi}^{k+1},\pi^{k+1})\\
\leq&|\Delta_{\widetilde{M}^{k+1}}(\widetilde{\pi}^{k+1},\pi^{k+1})-\Delta_{M^*}(\widetilde{\pi}^{k+1},\pi^{k+1})|+|\Delta_{M^*}(\widetilde{\pi}^{k+1},\pi^{k+1})-\Delta_{M^{k+1}}(\widetilde{\pi}^{k+1},\pi^{k+1})|.
\end{aligned}$$

where the first inequality is because of optimism, and the second one is because π^{k+1} is the equilibirum of M^{k+1}. Next, we provide a key lemma to upper bound the RHS.

Lemma 4.6. *Given two arbitrary model M and $\widetilde{M}$, and two policies π and $\widetilde{\pi}$, we have:*

$$\begin{aligned}
|\Delta_M(\widetilde{\pi},\pi)-\Delta_{\widetilde{M}}(\widetilde{\pi},\pi)|\leq&\mathbb{E}_{\widetilde{\pi},M|\boldsymbol{\mu}^\pi_M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]\\
&+(2L_rH+1)\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]. \qquad (9)
\end{aligned}$$

As we can see, to control the exploitability, we require the model can predict well on the data distribution induced by both π^{k+1} and $\widetilde{\pi}^{k+1}$ conditioning on $\boldsymbol{\mu}^{\pi^{k+1}}_{M^*}$, which motivates our formulation of Def. 2.2. By combining Lem. 4.6 and theorems in the first part, we finish the proof.

5 Exponential Separation between RL in MFC and MFG

In this section, we establish the separation between RL in MFC and MFG by investigating the sample complexity lower and upper bounds in tabular setting with function approximation. Our results are based on the generative model defined below, which is also frequently considered in previous Mean-Field [24, 20] or single-agent literatures [2]. We show that, under Assump. A and B, identifying ε-NE in MFG is exponentially more GM-efficient than finding ε-optimal policy in MFC.

Definition 5.1 (Generative Model)**.** The Generative Model (abbr. GM) can be queried by arbitrary $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h, \mu_h \in \Delta(\mathcal{S}_h)$, and return a sample from distribution $\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu_h)$.

5.1 Exponential Lower Bound in Tabular Mean-Field Control

In the following theorem, we establish an exponential sample complexity lower bound for tabular MFC (with function approximation) given access to both GM in Def. 5.1 and DCP in Def. 2.2, which indicates a separation with tabular single-agent RL. Intuitively, in the worst case, the agent should explore the entire $\mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$ space to identify the policy achieving the maximal return because of the dependence on μ in transition probabilities. We defer the proof for Thm. 5.2 to Appx. F.

Theorem 5.2. *[Exponential Lower Bound for MFC] Given arbitrary $L_T > 0$ and $d \geq 2$, 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 $\varepsilon \leq \frac{L_T}{d+1}$, 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 $K \leq |\mathcal{M}|/2 - 1$, the probability that Alg produces an ε-near-optimal policy is less than $1/2$.*

Our hard instance can be regarded as representation learning in low-rank MF-MDP generalized from single-agent setting [1, 42, 56]. In contrast, as we shown in Prop. 3.4, if the representation ϕ is known to the learner, the model class has low MBED regardless of the function class of ψ, and therefore, can be solved with polynomial samples by Alg. 1.

In single-agent RL, however, efficient learning is possible no matter the representation is available or not [35, 64, 1, 42, 56], although the sample complexity in representation learning has additional dependence on the number of actions $|\mathcal{A}|$ [1, 42, 56]. This separation becomes significant after generalizing to MFRL, where a "uniform cover" over the density space (similar role as the "uniform cover" over the action space) is required in the worst case.

5.2 Generative Model Efficient Learning in Tabular Mean-Field Game

In the previous section, we showed that MFC is sample-inefficient even when $H = 3$. In this section, we show that regardless of computational complexity, there exists an algorithm that can find an approximate NE policy consuming just polynomial samples from the GM. We state our algorithm in Alg. 2 and main result in Thm. 5.3, and defer its formal version and the proof to Appx. G.

Theorem 5.3 (GM-Efficiency; Informal)**.** *Under Assump. A and Assump. B, with appropriate hyper-parameter choices, w.p. at least $1 - \delta$, Alg. 2 returns an ε-approximate NE for M^* by consuming at most $O(\frac{S^3AH}{\varepsilon^2} \log^2 \frac{SAH|\mathcal{M}|}{\delta})$ queries to GM (dependence on Lipschitz factors are omitted).*

Algorithm Design and Highlight of Novelty We first introduce a new notation. Given a model class $\mathcal{M}$ and an arbitrary $M \in \mathcal{M}$ with transition function $\mathbb{P}_T$, given any $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h, \mu_h \in \Delta(\mathcal{S}_h)$, we use $\mathcal{B}^{\mathcal{M}}_{s_h,a_h,\mu_h}(M; \bar{\varepsilon})$ to denote the collection of models in $\mathcal{M}$, whose transition functions are $\bar{\varepsilon}$-close to M given s_h, a_h, μ_h. More concretely,

$$\mathcal{B}^{\mathcal{M}}_{s_h,a_h,\mu_h}(M; \bar{\varepsilon}) := \{\widetilde{M} \in \mathcal{M} | \|\mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu_h) - \mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h)\|_1 \leq \bar{\varepsilon}\}.$$

The basic idea of Alg. 2 is to find policies to gradually eliminate out inaccurate models in $\mathcal{M}$. In order to eliminate as much as possible inaccurate model, in the `If`-branch (line 3), we first try to find a (s_h, a_h, μ_h)-tuple, s.t., every model in $\mathcal{M}^k$ has significant number of models that disagree with it conditioning on (s_h, a_h, μ_h). In this case, by querying GM with $\widetilde{O}(\bar{\varepsilon}^2)$ samples, we can use $\mathbb{P}_{\widehat{T}^*}(\approx \mathbb{P}_{T^*})$ to eliminate models not in $\mathcal{B}^{\mathcal{M}^k}_{s_h,a_h,\mu_h}(M^*; \bar{\varepsilon})$, which is at least $\frac{1}{2}|\mathcal{M}^k|$.

The more challenging part is the `Else`-branch (line 7), when the models in $\mathcal{M}^k$ are not so easy to distinguish. Technically speaking, the separation between MFG and MFC is reflected by whether it is possible to efficiently eliminate models in this case. In the following Lemma, we provide a perspective to understand the tractability of MFG, where we show that the NE policy of model M is also the NE policy of M^* as long as they are "locally aligned" at that policy. In contrast, in MFC setting, such local alignment provides no information about whether the policy has optimal value.

Lemma 5.4. *[Implication of Local Alignment in MFG] Given a model M with transition $\mathbb{P}_T$, suppose M and M^* are locally aligned at policy π w.r.t. the density induced in M, i.e. $\forall h$, $\mathbb{P}_{T^*,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h}) = \mathbb{P}_{T,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h})$, if π is a NE in M, then it must be a NE in M^*.*

As our key technical novelty and contribution, in `Else`-branch, we propose Alg. 4 and prove that, Alg. 4 can construct a "Bridge Model" M^k_{Br} based on $\mathcal{M}^k$, such that (i) M^k_{Br} has at least one NE policy, denoted as $\pi^{\text{Br},k}_{\text{NE}}$; (ii) most of the models in $\mathcal{M}^k$ are (approximately) locally aligned at $\pi^{\text{Br},k}_{\text{NE}}$ w.r.t. the density $\boldsymbol{\mu}^{\text{NE},k}_{\text{Br}}$ induced in M^k_{Br}. As a result, either we can expect $\pi^{\text{Br},k}_{\text{NE}}$ is the approximate NE of M^*, or we can eliminate all the models locally aligned at this policy, which is at least $\frac{1}{2}|\mathcal{M}^k|$.

In summary, under good concentration events, either the loop continues but $|\mathcal{M}^{k+1}| \leq \frac{1}{2}|\mathcal{M}^k|$, or it returns an approximate NE. Therefore, we can conclude polynominal sample complexity to GM.

Algorithm 2: Equilibrium Finding by Model Elimination with Bridge Model

```
Input: Model Class $\mathcal{M}$, $\bar{\varepsilon}$, $\widetilde{\varepsilon}$, $K$, $N$, $\bar{N}$, $\widetilde{N}$.   Initialize: $\mathcal{M}^1 \leftarrow \mathcal{M}$.
for $k = 1, 2, ..., K$ do
    if $\exists (h, s_h, a_h, \mu_h)$, s.t. $\max_{M \in \mathcal{M}^k} |\mathcal{B}^{\mathcal{M}^k}_{s_h,\underline{a}_h,\mu_h}(M; \bar{\varepsilon})| \leq \frac{1}{2}|\mathcal{M}^k|$ then
        Query GM with $(h, s_h, a_h, \mu_h)$ for $\bar{N}$ samples, and compute empirical average as
         $\mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu_h)$.
        $\mathcal{M}^{k+1} \leftarrow \{M \in \mathcal{M}^k | \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu_h)\|_1 \leq \frac{\bar{\varepsilon}}{2}\}$
    end
    else
        $M^k_{\text{Br}}, \pi^{\text{Br},k}_{\text{NE}}, \boldsymbol{\mu}^{\text{NE},k}_{\text{Br}} \leftarrow$ BridgeModel$(\mathcal{M}^k, N)$. // Alg. 4 in Appx. G
        For any $(h, s_h, a_h)$, query GM with $(h, s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})$ for $\widetilde{N}$ samples, and compute
         empirical average as $\mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})$.
        if $\exists h, s_h, a_h$, s.t., $\|\mathbb{P}_{T^k_{Br},h}(\cdot|s_h, a_h, \mu^{NE,k}_{Br,h}) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu^{NE,k}_{Br,h})\|_1 \geq \widetilde{\varepsilon}$ then
            $\mathcal{M}^{k+1} \leftarrow \{M \in \mathcal{M}^k | \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1 \leq \frac{\widetilde{\varepsilon}}{2}\}$
        end
        else return $\pi^{\text{Br},k}_{\text{NE}}$ ;
    end
end
```

6 Conclusion and Open Problems

In this paper, we study the statistical efficiency of function approximation in MFRL. We propose the notion of MBED and an O-MLE based algorithm, which can guarantee to efficiently solve MFC and MFG given realizable function classes with bounded MBED. Besides, we provide evidence for the exponential separation between RL for MFC and MFG from an information-theoretic perspective. In the future, one important direction is to combine our results with optimization techniques to design computationally efficient algorithms. Moreover, it remains an open problem whether it is possible to extend our Alg. 2 and its guarantees to more general setting only with access to trajectory samples like Def. 2.2 instead of a Generative Model.

Acknowledgments and Disclosure of Funding

The authors would like to thank Andreas Krause for valuable discussion.

References

[1] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

[2] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

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

[4] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Lauriere. Reinforcement learning for mean field games, with applications to economics. *arXiv preprint arXiv:2106.13755*, 2021.

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

[6] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[7] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[8] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[9] Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

[10] Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

[11] Pierre Cardaliaguet and Charles-Albert Lehalle. Mean field game of controls and an application to trade crowding. *Mathematics and Financial Economics*, 12:335–363, 2018.

[12] René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

[13] Zixiang Chen, Chris Junchi Li, Angela Yuan, Quanquan Gu, and Michael I Jordan. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

[14] Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022.

[15] Areski Cousin, Stéphane Crépey, Olivier Guéant, David Hobson, Monique Jeanblanc, Jean-Michel Lasry, Jean-Paul Laurent, Pierre-Louis Lions, Peter Tankov, Olivier Guéant, et al. Mean field games and applications. *Paris-Princeton lectures on mathematical finance 2010*, pages 205–266, 2011.

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

[17] Qiwen Cui, Kaiqing Zhang, and Simon S Du. 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.

[18] Antonio De Paola, Vincenzo Trovato, David Angeli, and Goran Strbac. A mean field game approach for distributed control of thermostatic loads acting in simultaneous energy-frequency response markets. *IEEE Transactions on Smart Grid*, 10(6):5987–5999, 2019.

[19] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

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

[21] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[22] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

[23] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field multi-agent reinforcement learning: A decentralized network approach. *arXiv preprint arXiv:2108.02731*, 2021.

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

[25] Xin Guo, Renyuan Xu, and Thaleia Zariphopoulou. Entropy regularization for mean field games with learning. *Mathematics of Operations Research*, 2022.

[26] Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

[27] Jiawei Huang, Jinglin Chen, Li Zhao, Tao Qin, Nan Jiang, and Tie-Yan Liu. Towards deployment-efficient reinforcement learning: Lower bound and optimality. *arXiv preprint arXiv:2202.06450*, 2022.

[28] Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. 2006.

[29] Zool Hilmi Ismail, Nohaidda Sariff, and E Hurtado. A survey and analysis of cooperative multi-agent robot systems: challenges and directions. *Applications of Mobile Robots*, pages 8–14, 2018.

[30] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

[31] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

[32] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

[33] Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

[34] Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR, 2022.

[35] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

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

[37] Jae Won Lee, Jonghun Park, O Jangmin, Jongwoo Lee, and Euyseok Hong. A multiagent approach to q-learning for daily stock trading. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):864–877, 2007.

[38] Orin Levy, Asaf Cassel, Alon Cohen, and Yishay Mansour. Eluder-based regret for stochastic contextual mdps, 2022.

[39] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

[40] Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable markov games. *Advances in Neural Information Processing Systems*, 35:18296–18308, 2022.

[41] Aditya Mahajan. Reinforcement learning in stationary mean-field games. 2021.

[42] Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

[43] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[44] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[45] Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. *arXiv preprint arXiv:2210.16976*, 2022.

[46] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

[47] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends in Optimization*, 1(3):127–239, 2014.

[48] Barna Pasztor, Ilija Bogunovic, and Andreas Krause. Efficient model-based multi-agent mean-field reinforcement learning. *arXiv preprint arXiv:2107.04050*, 2021.

[49] Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up mean field games with online mirror descent. *arXiv preprint arXiv:2103.00623*, 2021.

[50] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

[51] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

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

[53] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[54] Shai Shalev-Shwartz and Yoram Singer. *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University, 2007.

[55] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

[56] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

[57] Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. *arXiv preprint arXiv:2302.06606*, 2023.

[58] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

[59] Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pages 11436–11447. PMLR, 2021.

[60] Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

[61] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang. A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning*, pages 24496–24523. PMLR, 2022.

[62] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR, 2018.

[63] Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. *arXiv preprint arXiv:2212.14449*, 2022.

[64] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[65] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[67] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2022.

Contents

A Extended Introduction

A.1 Motivation for Model-Based instead of Model-Free Algorithms

MFRL methods can be divided into two categories: model-based and model-free methods. Ours is the former, where we try to explicitly estimate the model. For the latter, it can be further divided into pure value-based (a value-function class is provided) and policy-based methods (a value function class and a policy class are provided). Although we can not assert model-based methods are the only solutions for MFRL, there are some special challenges for value-based methods that seem intractable, especially in function approximation setting.

For the pure value-based methods: In mean-field RL, the well-defined value function should be a function mapping from the joint space of states (or states, actions) and population density. Therefore, if one does not estimate the model, then it is not possible to track the density for each step and therefore, even if the true optimal value function is provided, one cannot convert it to the optimal policy since the density information is missing. In contrast, in single-agent setting, we can do that by simply taking argmax.

For the policy-based methods: Policy-based methods [24, 63], which repeatedly do the "policy improvement" step and "policy evaluation" step for the fixed policy, are not rare in MFRL in the tabular setting. However, one key difficulty is that this approach requires that the value function class can well approximate the value functions of all the possible policies encountered during the learning process. This assumption is feasible in the tabular setting because of finite states and actions, but it can be an extremely strong assumption in the non-tabular setting. Moreover, especially in MFG, strong assumptions, e.g. contractivity [24, 63], seems necessary to make sure the evaluated value function can provide useful information to optimize the policy to NE.

A.2 Other Related Work

Single-Agent RL with General Function Approximation Recently, in single agent setting, beyond tabular RL [6, 8, 31], there are significant progress on linear function approximation setting [35, 64, 1, 42, 56, 27] or more general conditions for efficient learning framework [51, 30, 55, 32, 19, 60, 21, 13, 67, 7] However, the MFRL setting is significantly different from single agent RL because of the dependence on density in transition and reward model. The function complexity measure, especially for value-based function class, and the corresponding algorithms in single-agent RL cannot be trivially generalized to MFRL.

Besides, the previous literatures discussing Model-Based Eluder Dimension is limited. Besides [38] which we compared with in Sec. 3, [46] considered the function class and their Eluder Dimension w.r.t. l_2 distance, which restricted the function classes it can capture.

Multi-Agent RL Sample complexity of learning in Markov Games has been studied extensively in a recent surge of works [33, 9, 14, 65, 66, 61]. A few recent works also consider learning Markov Games with linear or general function approximation [58, 26, 34, 45]. None of these results can be directly extended to Mean-Field RL.

Recently, [57, 17] also studied how to "break the curse of multi-agency" by decentralized learning in MARL setting. Although they consider a more general setting from ours, where they did not employ mean field assumption and allowed the agents can be largely different, there are still some restrictions when applying to our mean-field setting. First of all, their algorithm can only guarantee the convergence to the Coarse Correlated Equilibria or the Correlated Equilibria, while ours can converge to Nash Equilibrium in MFG. Moreover, and more importantly, their algorithms have sample complexity depending on the number of agents (although polynominal instead of exponential), which still suffer from the "curse of multi-agency" when the number of agents is exponentially large.

Other Related Works In this paper, we consider MLE based model estimation algorithm. Similar ideas has been adopted in POMDP [39] or Partial Observable Markov Games [40].

B Proofs for Eluder Dimension Related

B.1 Missing Details of Eluder Dimension Related

In the following, we recall the Eluder Dimension in Value Function Approximation Setting [51].

Definition B.1 (ε-Independence for Scalar Function)**.** Given a domain $\mathcal{Y}$ and a function class $\mathcal{F} \subset \{f|f : \mathcal{Y} \to \mathbb{R}\}$, we say y is ε-independent w.r.t. $y_1, y_2, ..., y_n$, if there exists $f_1, f_2 \in \mathcal{F}$ satisfying $\sqrt{\sum_{i=1}^n |f_1(y_i) - f_2(y_i)|^2} \leq \varepsilon$ but $|f_1(y) - f_2(y)| > \varepsilon$.

Definition B.2 (Eluder Dimension for Scalar Function)**.** Given a function class $\mathcal{F} \subset \{g|g : \mathcal{Y} \to \mathbb{R}\}$, we use $\overline{\text{dimE}}(\mathcal{F}, \varepsilon)$ to denote the length of the longest sequence $y_1, ..., y_n \in \mathcal{Y}$, such that, for any $i \in [n]$, y_i is ε-independent w.r.t. $y_1, ..., y_{i-1}$.

Remarks on Assump. B The main reason we require the Lipschitz continuity w.r.t. Hellinger distance is to handle the distribution shift issue. In Thm. 4.2, we show that MLE regression can only guarantee the learned model aligns with M^* under different density. In order to guarantee efficient learning, we need to convert it to upper bound for model error under the same density.

Besides, although in general $\mathbb{H}$ and $\mathbb{TV}$ distance between two distribution can be largely different. For our example in Prop. 3.5, given two function f_1, f_2, we have:

$$\mathbb{H}(\mathbb{P}_{f_1}, \mathbb{P}_{f_2}) = O(\frac{1}{\sigma^2}\|f_1 - f_2\|_2) = O(\frac{1}{\sigma^2}\|f_1 - f_2\|_1).$$

Therefore, Assump. B can be ensured when $f \in \mathcal{F}$ is Lipschitz w.r.t. l_1 distance, which is reasonable.

Moreover, in fact, if we only consider $\text{dimE}_\alpha(\mathcal{M}, \mathbb{TV}, \varepsilon)$ as model-based eluder dimension in our framework, we only require Lipschitz continuity w.r.t. l_1-distance (or $\mathbb{TV}$ distance).

B.2 Concrete Examples Satisfying Finite Eluder Dimension Assumption

B.2.1 Example 1: Linear Combined Model

Proposition B.3 (Linearly Combined Model)**.** *Consider the linear combined model class with known state action feature vector $\phi(s, a, \mu, s') \in \mathbb{R}^d$, such that for arbitrary $s \in \mathcal{S}, a \in \mathcal{A}$ and arbitrary $g : \mathcal{S} \to [0, 1]$, we have $\|\sum\!\!\!\!\!\!\int_{s' \in \mathcal{S}} \phi(s, a, \mu, s')g(s')\|_2 \leq C_\phi$*[3]

$$\mathcal{P} := \{\mathbb{P}_\theta|\mathbb{P}_\theta(\cdot|s, a, \mu) := \theta^\top \phi(s, a, \mu, s'),\ \|\theta\|_2 \leq C_\theta;\ \forall s, a, \mu,\ \sum\!\!\!\!\!\!\!\!\int_{s' \in \mathcal{S}} \mathbb{P}(s'|s, a, \mu) = 1,\ \mathbb{P}(\cdot|s, a, \mu) \geq 0\}.$$

For $\alpha \geq 1$, we have: $\text{dimE}_\alpha(\mathcal{P}, \mathbb{TV}, \varepsilon) = O(d \log(1 + \frac{dC_\theta C_\phi}{\varepsilon}))$.

Proof. We focus on the case when $\alpha = 1$ since which directly serves as upper bound for $\alpha > 1$. For arbitrary θ_1, θ_2 with $\|\theta_1\|_2 \leq C_\theta, \|\theta_2\|_2 \leq C_\theta$, we have:

$$\mathbb{TV}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2})(s, a, \mu) = \sup_{\bar{\mathcal{S}}} |\sum\!\!\!\!\!\!\!\!\int_{s'} (\theta_1 - \theta_2)^\top \phi(s, a, \mu, s')| = \frac{1}{2}(\theta_1 - \theta_2)^\top \sum\!\!\!\!\!\!\!\!\int_{s'} \phi(s, a, \mu, s')g_{\theta_1,\theta_2}(s, a, \mu, s').$$

where we define:

$$g_{\theta_1,\theta_2}(s, a, \mu, s') := \begin{cases} 1, & \text{if } (\theta_1 - \theta_2)^\top \phi(s, a, \mu, s') \geq 0; \\ -1, & \text{otherwise.} \end{cases}$$

In the following, for simplicity, we use

$$v_{\theta_1,\theta_2}(s, a, \mu) := \sum\!\!\!\!\!\!\!\!\int_{s'} \phi(s, a, \mu, s')g_{\theta_1,\theta_2}(s, a, \mu, s').$$

as a short note. Also note that,

$$\|v_{\theta_1,\theta_2}(s, a, \mu)\|_2 \leq \|\sum\!\!\!\!\!\!\!\!\int_{s'} \phi(s, a, \mu, s')g_{\theta_1,\theta_2}(s, a, \mu, s')\|_2 \leq C_\phi, \quad \forall \pi, \mu, \theta_1, \theta_2 \in \mathcal{B}(0; C_\theta).$$

[3] Similar normalization assumptions is common in previous literatures [1, 43, 56]

Suppose we have a sequence of samples $x_1, .., x_n$ with $x_i := (s^i, a^i, \mu^i)$, such that x_i is ε-independent of $\{x_1, ..., x_{i-1}\}$ for all $i \in [n]$. Then, by definition, for each i, there exists θ_1^i, θ_2^i such that:

$$\begin{aligned} 4\varepsilon^2 \leq & 4\|\mathbb{P}_{\theta_1^i}(\cdot|s^i, a^i, \mu^i) - \mathbb{P}_{\theta_2^i}(\cdot|s^i, a^i, \mu^i)\|_{\mathbb{TV}}^2 \\ = & \Big((\theta_1^i - \theta_2^i)^\top v_{\theta_1^i, \theta_2^i}(s^i, a^i, \mu^i)\Big)^2 \leq \|\theta_1^i - \theta_2^i\|_{\Lambda^i}^2 \|v_{\theta_1^i, \theta_2^i}(s^i, a^i, \mu^i)\|_{(\Lambda^i)^{-1}}^2. \end{aligned}$$

where we denote

$$\Lambda^i := \lambda I + \sum_{t=1}^{i-1} v_{\theta_1^t, \theta_2^t}(s^t, a^t, \mu^t)^\top v_{\theta_1^t, \theta_2^t}(s^t, a^t, \mu^t).$$

Meanwhile,

$$\begin{aligned} \|\theta_1^i - \theta_2^i\|_{\Lambda^i}^2 = & \lambda\|\theta_1^i - \theta_2^i\|^2 + \sum_{t=1}^{i-1} \Big((\theta_1^i - \theta_2^i)^\top v_{\theta_1^t, \theta_2^t}(s^t, a^t, \mu^t)\Big)^2 \\ \leq & 4\lambda C_\theta^2 + \sum_{t=1}^{i-1} \Big((\theta_1^i - \theta_2^i) \mathop{\sum\!\!\!\!\!\!\int}_{s'} \Phi(s^t, a^t, \mu^t)\psi(s') g_{\theta_1^t, \theta_2^t}(s^t, a^t, \mu^t, s')\Big)^2 \\ = & 4\lambda C_\theta^2 + 4\sum_{t=1}^{i-1} \|\mathbb{P}_{\theta_1^t}(\cdot|s^t, a^t, \mu^t), \mathbb{P}_{\theta_2^t}(\cdot|s^t, a^t, \mu^t)\|_{\mathbb{TV}} \qquad (|g_{\theta_1^t, \theta_2^t}(\cdot, \cdot, \cdot, \cdot)| = 1) \\ \leq & 4\lambda C_\theta^2 + 4\varepsilon^2. \end{aligned}$$

By choosing $\lambda = \varepsilon^2 / C_\theta^2$, we further have:

$$\|v_{\theta_1^i, \theta_2^i}(s^i, a^i, \mu^i)\|_{(\Lambda^i)^{-1}}^2 \geq \frac{4\varepsilon^2}{4\lambda C_\theta^2 + 4\varepsilon^2} = \frac{1}{2}.$$

On the one hand,

$$\begin{aligned} \det \Lambda^{n+1} = & \det(\Lambda^n + v_{\theta_1^n, \theta_2^n}(s^n, a^n, \mu^n) v_{\theta_1^n, \theta_2^n}(s^n, a^n, \mu^n)^\top) \\ = & (1 + v_{\theta_1^n, \theta_2^n}(s^n, a^n, \mu^n)^\top (\Lambda^n)^{-1} v_{\theta_1^n, \theta_2^n}(s^n, a^n, \mu^n)) \cdot \det \Lambda^n \\ \geq & \frac{3}{2} \det \Lambda^n \geq (\frac{3}{2})^n \det \Lambda^1 = \lambda^d (\frac{3}{2})^n. \end{aligned}$$

On the other hand,

$$\lambda^d (\frac{3}{2})^n \leq \det \Lambda^{n+1} \leq (\frac{\text{Tr}(\Lambda^n)}{d})^d \leq (\lambda + \frac{n C_\phi^2}{d})^d.$$

which implies $n = O(d \log(1 + \frac{d C_\theta C_\phi}{\varepsilon}))$. □

Linear Combined Model with State-Action-Dependent Weights In [43], the authors introduced another style of linear combined model with state-action dependent weights, which can be generalized to MFRL setting by:

$$\mathbb{P}_W(s'|s, a, \mu) := \sum_{i=1}^{d} [W\phi(s, a, \mu, s')]_k \mathbb{P}_i(s'|s, a, \mu).$$

where $W \in \mathbb{R}^{d \times d}$ is an unknown matrix, $\phi(s, a)$ are known feature class, $\{\mathbb{P}_i\}_{i=1}^d$ are d known models to combine. If we further define $\psi(s, a, \mu, s') := [\mathbb{P}_1(s'|s, a, \mu), ..., \mathbb{P}_d(s'|s, a, \mu)]^\top \in \mathbb{R}^d$, we can rewrite the model by:

$$\mathbb{P}_W(s'|s, a, \mu) = \phi(s, a, \mu, s')^\top W^\top \psi(s, a, \mu, s') = \mathbf{vec}(W^\top)^\top \mathbf{vec}(\psi(s, a, \mu, s')\phi(s, a, \mu, s')^\top).$$

Therefore, by treating $\theta = \mathbf{vec}(W^\top)$ to be the parameter and $\mathbf{vec}(\psi(s, a, \mu, s')\phi(s, a, \mu, s')^\top)$ to be the feature taking place the role of $\phi(s, a, \mu, s')$ in Prop. B.3, we can absorb this model class into linearly combined model framework, and $\widetilde{O}(d^2)$ will be an upper bound for its MBED.

B.2.2 Example 2: Linear MDP with Known Feature

Proposition B.4 (Low-Rank MF-MDP; Formal Version of Prop. 3.4)**.** *Consider the Low-Rank MF-MDP with known feature $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \rightarrow \mathbb{R}^d$ satisfying $\|\phi\| \leq C_\phi$, and unknown next state feature $\psi : \mathcal{S} \rightarrow \mathbb{R}^d$. Given a next state feature function class Ψ satisfying $\forall \psi \in \Psi$, $\forall s' \in \mathcal{S}$, $\forall g : \mathcal{S} \rightarrow \{-1, 1\}$, $\| \mathop{\sum\!\!\!\!\!\!\int}_{s'} \psi(s')g(s')\|_2 \leq C_\Psi$, consider the following model class:*

$$\mathcal{P}_\Psi := \{\mathbb{P}_\psi | \mathbb{P}_\psi(\cdot|s, a, \mu) := \phi(s, a, \mu)^\top \psi(s'); \forall s, a, \mu, \ \mathop{\sum\!\!\!\!\!\!\int}_{s' \in \mathcal{S}} \mathbb{P}_\psi(s'|s, a, \mu) = 1, \ \mathbb{P}_\psi(s'|s, a, \mu) \geq 0; \psi \in \Psi\},$$

for $\alpha \geq 1$, we have $\dim E_\alpha(\mathcal{P}_\Psi, \mathbb{TV}, \varepsilon) = O(d \log(1 + \frac{dC_\phi C_\Psi}{\varepsilon}))$.

Proof. Again we focus on the case when $\alpha = 1$. Suppose there is a sequence of samples $x_1, ..., x_n$ (with $x_i := (s^i, a^i, \mu^i)$) such that for any $i \in [n]$, x_i is ε-independent w.r.t. $x_1, ..., x_{i-1}$ w.r.t. $\mathcal{P}_\Psi$ and $\mathbb{TV}$, then for each $i \in [n]$, there should exists $\psi^i, \widetilde{\psi}^i \in \Psi$, such that:

$$\varepsilon^2 \geq \sum_{t=1}^{i-1} \|\mathbb{P}_{\psi^i}(\cdot|s^t, a^t, \mu^t), \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s^t, a^t, \mu^t)\|_{\mathbb{TV}}^2.$$

and

$$\begin{aligned}
\varepsilon^2 \leq & \|\mathbb{P}_{\psi^i}(\cdot|s^i, a^i, \mu^i) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s^i, a^i, \mu^i)\|_{\mathbb{TV}}^2 \\
= & \sup_{\bar{\mathcal{S}} \subset \mathcal{S}} \Big(\mathop{\sum\!\!\!\!\!\!\int}_{s' \in \bar{\mathcal{S}}} \phi(s^i, a^i, \mu^i)^\top (\psi^i(s') - \widetilde{\psi}^i(s')) \Big)^2 \\
= & \frac{1}{4} \Big(\phi(s^i, a^i, \mu^i)^\top \mathop{\sum\!\!\!\!\!\!\int}_{s' \in \mathcal{S}} (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i, s') \Big)^2 \\
\leq & \frac{1}{4} \|\phi(s^i, a^i, \mu^i)\|_{(\Lambda^i)^{-1}}^2 \| \mathop{\sum\!\!\!\!\!\!\int}_{s' \in \mathcal{S}} (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i, s') \|_{\Lambda^i}^2.
\end{aligned}$$

where we define:

$$\Lambda^i := \lambda I + \sum_{t=1}^{i-1} \phi(s^i, a^i, \mu^i) \phi(s^i, a^i, \mu^i)^\top; \quad g_{\psi^i, \widetilde{\psi}^i}(s, a, \mu, s') := \begin{cases} 1, & \text{if } \phi(s^i, a^i, \mu^i)^\top (\psi^i(s') - \widetilde{\psi}^i(s')) \geq 0; \\ -1, & \text{otherwise.} \end{cases}$$

For simplicity, we use $v_{\psi, \widetilde{\psi}}(s, a, \mu) := \mathop{\sum\!\!\!\!\!\!\int}_{s'} (\psi(s') - \widetilde{\psi}(s')) g_{\psi, \widetilde{\psi}}(s, a, \mu, s')$ as a shortnote. Therefore, for each i,

$$\begin{aligned}
\|v_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i)\|_{\Lambda^i}^2 = & \lambda \|v_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i)\|^2 + \sum_{t=1}^{i-1} \Big(\phi(s^t, a^t, \mu^t)^\top v_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i) \Big)^2 \\
= & \lambda \|v_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i)\|^2 + \sum_{t=1}^{i-1} \Big(\phi(s^t, a^t, \mu^t)^\top \mathop{\sum\!\!\!\!\!\!\int}_{s'} (\psi^i(s) - \widetilde{\psi}^i(s')) g_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i, s') \Big)^2 \\
= & 4\lambda C_\Psi^2 + 4 \sum_{t=1}^{i-1} \|\mathbb{P}_{\psi^i}(\cdot|s^t, a^t, \mu^t) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s^t, a^t, \mu^t)\|_{\mathbb{TV}}^2 \\
\leq & 4\lambda C_\Psi^2 + 4\varepsilon^2.
\end{aligned}$$

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

$$\|\phi(s^i, a^i, \mu^i)\|_{(\Lambda^i)^{-1}}^2 \geq \frac{4\varepsilon^2}{4\lambda C_\Psi^2 + 4\varepsilon^2} = \frac{1}{2}.$$

On the one hand,

$$\det \Lambda^{n+1} = \det(\Lambda^n + \phi(s^n, a^n, \mu^n)\phi(s^n, a^n, \mu^n)^\top) = (1 + \|\phi(s^n, a^n, \mu^n)\|_{(\Lambda^n)^{-1}}^2) \cdot \det \Lambda^n$$

$$\geq \frac{3}{2}\det \Lambda^n \geq (\frac{3}{2})^n \det \Lambda^1 = \lambda^d (\frac{3}{2})^n.$$

Therefore,

$$\lambda^d (\frac{3}{2})^n \leq \det \Lambda^{n+1} \leq (\frac{\text{Tr}(\Lambda^n)}{d})^d \leq (\lambda + \frac{nC_\phi^2}{d})^d.$$

which implies $n = O(d \log(1 + \frac{dC_\phi C_\Psi}{\varepsilon}))$. □

B.2.3 Example 3: Kernel Mean-Field MDP

We first introduce the notion of Effective Dimension, which is also known as the critical information gain in [19]:

Definition B.5 (Effective Dimension)**.** The ε-effective dimension of a set $\mathcal{Y}$ is the minimum integer $d_{\text{eff}}(\mathcal{Y}, \varepsilon) = n$, such that,

$$\sup_{y_1,...,y_n \in \mathcal{Y}} \frac{1}{n} \log \det(I + \frac{1}{\varepsilon^2} \sum_{i=1}^n y_i y_i^\top) \leq \frac{1}{e}.$$

In the next theorem, we show that, the MBED of kernel MF-MDP generalized from kernel MDP in single-agent setting [32] can be upper bounded by the effiective dimension in certain Hilbert spaces.

Proposition B.6 (Kernel MF-MDP)**.** *Given a separable Hilbert space $\mathcal{H}$, a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \rightarrow \mathcal{H}$ such that $\|\phi(s, a, \mu)\|_\mathcal{H} \leq C_\phi$ for all $s \in \mathcal{S}, a \in \mathcal{A}, \mu \in \Delta(\mathcal{S})$, and a next state feature class $\Psi \subset \{\psi : \mathcal{S} \rightarrow \mathcal{H}\}$ satisfying the normalization property that $\forall\ \psi \in \Psi$ and $g : \mathcal{S} \rightarrow \{-1, 1\}$, $\| \sumint_{s' \in \mathcal{S}} \psi(s') g(s') \|_\mathcal{H} \leq 1$* [4]*. Consider the model class $\mathcal{P}_{\Psi,\mathcal{H}}$ defined by:*

$$\mathcal{P}_{\Psi,\mathcal{H}} := \{\mathbb{P}_\psi | \mathbb{P}_\psi(s'|s, a, \mu) = \langle \phi(s, a, \mu), \psi(s') \rangle_\mathcal{H},\ \sumint_{s' \in \mathcal{S}} \mathbb{P}_\psi(s'|s, a, \mu) = 1,\ \mathbb{P}_\psi(\cdot|s, a, \mu) \geq 0,\ \psi \in \Psi\}.$$

For $\alpha \geq 1$, we have

$$\mathit{dimE}_\alpha(\mathcal{P}_{\Psi,\mathcal{H}}, \mathbb{TV}, \varepsilon) = O(d_{\mathit{eff}}(\phi(\mathcal{X}), \varepsilon)),$$

where we use $\mathcal{X} := \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$ as a short note, and $\phi(\mathcal{X}) := \{\phi(x) | x \in \mathcal{X}\}$.

Proof. The proof idea is similar to the proof of Prop. B.4. Again, we only focus on the case when $\alpha = 1$. Suppose there is a sequence of samples $x_1, ..., x_n$ (with $x_i := (s^i, a^i, \mu^i)$) such that for any $i \in [n]$, x_i is ε-independent w.r.t. $x_1, ..., x_{i-1}$ w.r.t. $\mathcal{P}_{\Psi,\mathcal{H}}$ and $\mathbb{TV}$, then for each $i \in [n]$, there should exists $\psi^i, \widetilde{\psi}^i \in \Psi$, such that:

$$\varepsilon^2 \geq \sum_{t=1}^{i-1} \|\mathbb{P}_{\psi^i}(\cdot|s^t, a^t, \mu^t) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s^t, a^t, \mu^t)\|_{\mathbb{TV}}^2.$$

and

$$\begin{aligned}
4\varepsilon^2 \leq & 4\|\mathbb{P}_{\psi^i}(\cdot|s^i, a^i, \mu^i) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s^i, a^i, \mu^i)\|_{\mathbb{TV}}^2 \\
= & \Big(\langle \phi(s^i, a^i, \mu^i), \sumint_{s'} (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i, s') \rangle_\mathcal{H}\Big)^2 \\
\leq & \|\phi(s^i, a^i, \mu^i)\|_{(\Lambda^i)^{-1}}^2 \| \sumint_{s'} (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i, s') \|_{\Lambda^i}^2.
\end{aligned}$$

where we define:

$$\Lambda^i := \lambda I + \sum_{t=1}^{i-1} \phi(s^i, a^i, \mu^i)\phi(s^i, a^i, \mu^i)^\top; \quad g_{\psi^i, \widetilde{\psi}^i}(s, a, \mu, s') := \begin{cases} 1, & \text{if } \langle \phi(s^i, a^i, \mu^i), \psi^i(s') - \widetilde{\psi}^i(s') \rangle_\mathcal{H} \geq 0; \\ -1, & \text{otherwise.} \end{cases}$$

[4] To align with [32], we assume ψ is normalized.

Based on a similar discussion and choice of $\lambda = \varepsilon^2$, as Prop. B.4, we have:

$$(\frac{3}{2})^n \det \Lambda^1 \leq \det \Lambda^{n+1} = \det(\varepsilon^2 I + \sum_{i=1}^n \phi(s^i, a^i, \mu^i)\phi(s^i, a^i, \mu^i)^\top),$$

Therefore,

$$n \log \frac{3}{2} \leq \frac{\det \Lambda^{n+1}}{\det \Lambda^1} = \det(I + \frac{1}{\varepsilon^2} \sum_{i=1}^n \phi(s^i, a^i, \mu^i)\phi(s^i, a^i, \mu^i)^\top) \leq \frac{1}{e} d_{\text{eff}}(\phi(\mathcal{X}), \varepsilon),$$

which implies $n = O(d_{\text{eff}}(\phi(\mathcal{X}), \varepsilon))$. □

B.2.4 Example 4: Generalized Linear Function Class

In this section, we extend the Generalized Linear Models in single-agent RL [51] to MF-MDP.

Proposition B.7 (Generalized Linear MF-MDP)**.** *Given a differentiable and strictly increasing function $h : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $0 < \underline{h} \leq h' \leq \overline{h}$, where h' is its derivative, suppose we have a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \rightarrow \mathbb{R}^d$ satisfying $\|\phi(\cdot, \cdot, \cdot)\|_2 \leq C_\phi$ and a feature class $\Psi \subset \{\psi | \psi : \mathcal{S} \rightarrow \mathbb{R}\}$ such that for any $\psi \in \Psi$, $\| \sumint_{s' \in \mathcal{S}} \psi(s') g(s') \|_2 \leq C_\Psi$ for any $g : \mathcal{S} \rightarrow \{-1, 1\}$. Consider the model class:*

$$\mathcal{P}_{h,\Psi} := \{\mathbb{P}_\psi | \mathbb{P}_\psi(\cdot|s, a, \mu) := h(\phi(s, a, \mu)^\top \psi(s')); \forall s, a, \mu,\ \|\mathbb{P}_\psi(\cdot|s, a, \mu)\|_1 = 1,\ \mathbb{P}_\psi(s'|s, a, \mu) \geq 0; \psi \in \Psi\},$$

For any $\alpha \geq 1$, we have dimE$_\alpha(\mathcal{P}_{h,\Psi}, \mathbb{TV}, \varepsilon) = \widetilde{O}(dr^2)$, where $r := \overline{h}/\underline{h}$.

Proof. The proof is similar to Prop. B.4. Suppose there is a sequence of samples $x_1, ..., x_n$ (with $x_i := (s^i, a^i, \mu^i)$) such that for any $i \in [n]$, x_i is ε-independent w.r.t. $x_1, ..., x_{i-1}$ w.r.t. $\mathcal{P}_\Psi$ and $\mathbb{TV}$, then for each $i \in [n]$, there should exists $\psi^i, \widetilde{\psi}^i \in \Psi$, such that:

$$\begin{aligned}
\varepsilon^2 \geq & \sum_{t=1}^{i-1} \|\mathbb{P}_{\psi^i}(\cdot|s^t, a^t, \mu^t) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s^t, a^t, \mu^t)\|_{\mathbb{TV}}^2 \\
= & \sum_{t=1}^{i-1} \sup_{\bar{\mathcal{S}} \subset \mathcal{S}} \Big(\sumint_{s' \in \bar{\mathcal{S}}} h(\phi(s^t, a^t, \mu^t)^\top \psi^i(s')) - h(\phi(s^t, a^t, \mu^t)^\top \widetilde{\psi}^i(s')) \Big)^2 \\
\geq & \underline{h}^2 \sum_{t=1}^{i-1} \sup_{\bar{\mathcal{S}} \subset \mathcal{S}} \Big(\sumint_{s' \in \bar{\mathcal{S}}} \phi(s^t, a^t, \mu^t)^\top (\psi^i(s') - \widetilde{\psi}^i(s')) \Big)^2. \qquad \text{(Mean Value Theorem)}
\end{aligned}$$

Besides,

$$\begin{aligned}
4\varepsilon^2 \leq & 4\|\mathbb{P}_{\psi^i}(\cdot|s^i, a^i, \mu^i) - \mathbb{P}_{\widetilde{\psi}^i}(\cdot|s^i, a^i, \mu^i)\|_{\mathbb{TV}}^2 \\
= & 4 \sup_{\bar{\mathcal{S}} \subset \mathcal{S}} \Big(\sumint_{s' \in \bar{\mathcal{S}}} h(\phi(s^i, a^i, \mu^i)^\top \psi^i(s')) - h(\phi(s^i, a^i, \mu^i)^\top \widetilde{\psi}^i(s')) \Big)^2 \\
\leq & 4\overline{h}^2 \sup_{\bar{\mathcal{S}} \subset \mathcal{S}} \Big(\sumint_{s' \in \bar{\mathcal{S}}} \phi(s^i, a^i, \mu^i)^\top (\psi^i(s') - \widetilde{\psi}^i(s')) \Big)^2 \\
= & \overline{h}^2 \Big(\sumint_{s' \in \bar{\mathcal{S}}} \phi(s^i, a^i, \mu^i)^\top (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i, s') \Big)^2 \\
\leq & \overline{h}^2 \|\phi(s^i, a^i, \mu^i)\|_{(\Lambda^i)^{-1}}^2 \| \sumint_{s'} (\psi^i(s') - \widetilde{\psi}^i(s')) g_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i, s') \|_{\Lambda^i}^2.
\end{aligned}$$

where in the second inequality, we use the mean value theorem and the fact that $h' \leq \overline{h}$; Λ^i and $g_{\psi^i, \widetilde{\psi}^i}$ are the same as those in Prop. B.4. By denoting $v_{\psi, \widetilde{\psi}}(s, a, \mu) := \sumint_{s'} (\psi(s') - \widetilde{\psi}(s')) g_{\psi, \widetilde{\psi}}(s, a, \mu, s')$, similar to the proof in Prop. B.4, we have the following upper bound:

$$\|v_{\psi^i, \widetilde{\psi}^i}(s^i, a^i, \mu^i)\|_{\Lambda^i}^2 \leq 4\lambda C_\Psi^2 + 4\varepsilon^2 / \underline{h}^2.$$

By choosing $\lambda = \varepsilon^2/\underline{h}^2 C_\Psi^2$, we have:

$$\|\phi(s^i, a^i, \mu^i)\|^2_{(\Lambda^i)^{-1}} \geq \frac{4\varepsilon^2}{\overline{h}(4\lambda C_\Psi^2 + 4\varepsilon^2/\underline{h}^2)} = \frac{1}{r^2}.$$

By a similar discussion, we have:

$$(1 + \frac{1}{r^2})^n \det \Lambda^1 \leq \det \Lambda^{n+1} \leq (\lambda + \frac{nC_\phi^2}{d})^d.$$

which implies:

$$n = O(d \log(1 + \frac{\overline{h} d C_\phi C_\Psi}{\varepsilon}) / \log(1 + \frac{1}{r^2})) = O(dr^2 \log(1 + \frac{\overline{h} d C_\phi C_\Psi}{\varepsilon})).$$

□

B.2.5 Example 3: Deterministic Transition with Gaussian Noise

Proposition 3.5. *[Deterministic Transition with Gaussian Noise] Suppose $\mathcal{S} \subset \mathbb{R}^d$. Given a function class $\mathcal{G} \subset \{g | g : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \times \mathbb{N}^* \to \mathbb{R}\}$ and convert it to $\mathcal{F}_\mathcal{G} := \{f_g | f_g(\cdot,\cdot,\cdot) := [g(\cdot,\cdot,\cdot,1), ..., g(\cdot,\cdot,\cdot,d)]^\top \in \mathbb{R}^d,\ g \in \mathcal{G}\}$. Consider the model class $\mathcal{P}_\mathcal{G} := \{\mathbb{P}_f | \mathbb{P}_f(\cdot|s,a,\mu) \sim f(s,a,\mu) + \mathcal{N}(0,\Sigma), f \in \mathcal{F}_\mathcal{G}\}$, where $\mathcal{N}(0,\Sigma)$ is the Gaussian noise with $\Sigma := \text{Diag}(\sigma, ..., \sigma)$. For $\varepsilon \leq 0.3$, we have $dimE_{\sqrt{2}}(\mathcal{P}_\mathcal{G}, \mathbb{H}, \varepsilon) \leq \overline{dimE}(\mathcal{F}_\mathcal{G}, 4\sigma\varepsilon)$, $dimE_{\sqrt{2d}}(\mathcal{P}_\mathcal{G}, \mathbb{H}, \varepsilon) \leq \overline{dimE}(\mathcal{G}, 4\sigma\varepsilon)$, where $\overline{dimE}$ is the Eluder Dimension for scalar or vector-valued functions [51, 46].*

Proof. First of all, consider the function $h(x) = 1 - \exp(-x/8)$, in general, we have:

$$\frac{x}{8} \geq h(x).$$

Besides, for $x \in [0,1]$, we have $0 \leq h(x) \leq 1 - \exp(-1/8)$ and

$$h(x) = 1 - \exp(-\frac{x}{8}) = \exp(0) - \exp(-\frac{x}{8}) \geq -\frac{\exp(-\frac{1}{8}) - \exp(0)}{1 - 0} x > \frac{1}{16} x.$$

Given $\varepsilon \leq 0.3 < \sqrt{1 - \exp(-1/8)}$, suppose we have a sequence of samples $x_1, ..., x_n \in \mathcal{X} := \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$, with $x_i := (s^i, a^i, \mu^i)$, such that for any $i \in [n]$, x_i is α-weakly-ε-independent w.r.t. $x_1, ..., x_{i-1}$. For any $i \in [n]$, there must exists $g_i^1, g_i^2 \in \mathcal{G}$ such that, $f_{g_i^1}, f_{g_i^2} \in \mathcal{F}_\mathcal{G}$, and

$$\begin{aligned}
\varepsilon^2 \geq & \sum_{j=1}^{i-1} \mathbb{H}^2(\mathbb{P}_{f_{g_i^1}}(\cdot|s^j, a^j, \mu^j), \mathbb{P}_{f_{g_i^2}}(\cdot|s^j, a^j, \mu^j)) \\
= & \sum_{j=1}^{i-1} h(\|f_{g_i^1}(s^j, a^j, \mu^j) - f_{g_i^2}(s^j, a^j, \mu^j)\|^2_{\Sigma^{-1}}) \\
\geq & \sum_{j=1}^{i-1} \frac{1}{16\sigma^2} \|f_{g_i^1}(s^j, a^j, \mu^j) - f_{g_i^2}(s^j, a^j, \mu^j)\|_2^2. \\
= & \frac{1}{16\sigma^2} \sum_{j=1}^{i-1} \sum_{t=1}^{d} |g_i^1(s^j, a^j, \mu^j, t) - g_i^2(s^j, a^j, \mu^j, t)|^2.
\end{aligned}$$

and

$$\begin{aligned}
\alpha^2\varepsilon^2 < & \mathbb{H}^2(\mathbb{P}_{f_{g_i^1}}(\cdot|s^i, a^i, \mu^i), \mathbb{P}_{f_{g_i^2}}(\cdot|s^i, a^i, \mu^i)) \\
\leq & \frac{1}{8\sigma^2} \|f_{g_i^1}(s^i, a^i, \mu^i) - f_{g_i^2}(s^i, a^i, \mu^i)\|_2^2 \\
= & \frac{1}{8\sigma^2} \sum_{t=1}^{d} |g_i^1(s^i, a^i, \mu^i, t) - g_i^2(s^i, a^i, \mu^i, t)|^2
\end{aligned}$$

$$\leq \frac{d}{8\sigma^2} \max_{t\in[d]} |g_i^1(s^i, a^i, \mu^i, t) - g_i^2(s^i, a^i, \mu^i, t)|^2.$$

By choosing $\alpha = \sqrt{2}$, we know that, for any $i \in [n]$, x_i is $4\sigma\varepsilon$-independent w.r.t. $\{x_1, ..., x_{i-1}\}$ on function class $\mathcal{F}_\mathcal{G}$. Therefore,

$$\text{dimE}_{\alpha=\sqrt{2}}(\mathcal{P}_\mathcal{G}, \mathbb{H}, \varepsilon) \leq \text{dimE}_{\alpha=1}(\mathcal{F}_\mathcal{G}, 4\sigma\varepsilon).$$

Besides, considering the sequence $t_1, t_2, ..., t_n$ with

$$t_i := \arg\max_{t\in[d]} |g_i^1(s^i, a^i, \mu^i, t) - g_i^2(s^i, a^i, \mu^i, t)|^2,$$

and choosing $\alpha = \sqrt{2d}$, we have (s^i, a^i, μ^i, t_i) is $4\sigma\varepsilon$-independent w.r.t. $\{(s^1, a^1, \mu^1, t_1), ..., (s^{i-1}, a^{i-1}, \mu^{i-1}, t_{i-1})\}$ for any $i \in [n]$. Therefore,

$$\text{dimE}_{\alpha=\sqrt{2d}}(\mathcal{P}_\mathcal{G}, \mathbb{H}, \varepsilon) \leq \overline{\text{dimE}}(\mathcal{G}, 4\sigma\varepsilon).$$

□

B.3 From Eluder Dimension to Regret Bound

Lemma B.8. *Under the condition and notation as Def. 3.1, consider a fixed $f^* \in \mathcal{F}$, and suppose we have a sequence $\{f_k\}_{k=1}^K \in \mathcal{F}$ and $\{x_k\}_{k=1}^K$ with $x_k := (s^k, a^k, \mu^k) \in \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$ satisfying that, for all $k \in [K]$, $\sum_{i=1}^{k-1} \boldsymbol{D}^2(f_k, f^*)(x_i) \leq \beta$. Then for all $k \in [K]$, and arbitrary $\varepsilon > 0$, we have:*

$$\sum_{k=1}^K \mathbb{I}[\boldsymbol{D}(f_k, f^*)(x_k) > \alpha\varepsilon] \leq (\frac{\beta}{\varepsilon^2} + 1)\textit{dimE}_\alpha(\mathcal{F}, \varepsilon).$$

Proof. We first show that, for some k, if $\mathbf{D}(f_k, f^*)(x_k) > \alpha\varepsilon$, then x_k is ε-dependent on at most β/ε^2 disjoint sub-sequence in $\{x_1, ..., x_{k-1}\}$. To see this, by Def. 3.3, if $\mathbf{D}(f_k, f^*)(x_k) > \alpha\varepsilon$ and x_k is α-weakly-ε-dependent w.r.t. a sub-sequence $\{x_{k_1}, ..., x_{k_\kappa}\} \subset \{x_i\}_{i=1}^{k-1}$, we must have:

$$\sum_{i=1}^\kappa \mathbf{D}^2(f_k, f^*)(x_{k_i}) \geq \varepsilon^2.$$

Given that $\sum_{i=1}^{k-1} \mathbf{D}^2(f_k, f^*)(x_i) \leq \beta$, the number of such kind of disjoint sub-sequence is upper bounded by β/ε^2.

On the other hand, for arbitrary sub-sequence $\{x_{k_1}, ..., x_{k_\kappa}\} \subset \{x_i\}_{i=1}^{k-1}$, there exists $j \in [\kappa]$ such that x_{k_j} is α-weakly-ε-dependent on $L := \lfloor \kappa/\text{dimE}_\alpha(\mathcal{F}, \varepsilon) \rfloor$ disjoint sub-sequence of $\{x_{k_1}, ...x_{k_{j-1}}\}$. To see this, we first construct L bins $B_1 = \{x_{k_1}\}, ..., B_L = \{x_{k_L}\}$. Then, we start with $j = L + 1$, and if x_{k_j} is already α-weakly-ε-dependent w.r.t. sequences $B_1, ..., B_L$, then we finish directly. Otherwise, there must exists B_l for some $l \in [L]$ such that x_{k_j} is α-weakly-ε-independent w.r.t. B_l, and we set $B_l \leftarrow B_l \cup \{x_{k_j}\}$ and $j \leftarrow j + 1$. Because the MBED is bounded, B_l can not be larger than $\text{dimE}_\alpha(\mathcal{F}, \varepsilon)$ if the above process continues. Therefore, the process must stop before $j \leq L \cdot \text{dimE}_\alpha(\mathcal{F}, \varepsilon) \leq \kappa$.

For arbitrary fixed $k \in [K]$, we use $\{x_{k_1}, ..., x_{k_\kappa}\} \subset \{x_1, ..., x_{k-1}\}$ to denote the elements such that $\mathbf{D}(f_i, f^*)(x_{k_i}) > \alpha\varepsilon$ for $i \in [\kappa]$. There must exists $j \in [\kappa]$, such that, on the one hand, x_{k_j} is α-weakly-ε-dependent with at most β/ε^2 disjoint sub-sequence of $\{x_{k_1}, ...x_{k_{j-1}}\}$, and on the other hand, x_{k_j} is α-weakly-ε-dependent on at least $L := \lfloor \kappa/\text{dimE}_\alpha(\mathcal{F}, \varepsilon) \rfloor$ disjoint sub-sequence of $\{x_{k_1}, ...x_{k_{j-1}}\}$. Therefore, we have:

$$\frac{\beta}{\varepsilon^2} \geq \lfloor \kappa/\text{dimE}_\alpha(\mathcal{F}, \varepsilon) \rfloor \geq \kappa/\text{dimE}_\alpha(\mathcal{F}, \varepsilon) - 1.$$

which implies $\kappa \leq (\frac{\beta}{\varepsilon^2} + 1)\text{dimE}_\alpha(\mathcal{F}, \varepsilon)$. □

Lemma 4.4. *Under the condition as Def. 3.1, consider a fixed $f^* \in \mathcal{F}$, and suppose we have a sequence $\{f_k\}_{k=1}^K \in \mathcal{F}$ and $\{x_k\}_{k=1}^K \subset \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$, s.t., for all $k \in [K]$, $\sum_{i=1}^{k-1} \boldsymbol{D}^2(f_k, f^*)(x_i) \leq \beta$, then for any $\varepsilon > 0$, we have $\sum_{k=1}^K \boldsymbol{D}(f_k, f^*)(x_k) = O(\sqrt{\beta K \textit{dimE}_\alpha(\mathcal{M}, \varepsilon)} + \alpha K\varepsilon)$.*

Proof. We first sort the sequence $\{\mathbf{D}(f_k, f^*)(x_k)\}_{k=1}^K$ and denote them by $e_1, e_2, ..., e_k$ with $e_1 \geq e_2 ... \geq e_K$. For $t \in [K]$, given any $\varepsilon > 0$, by Lem. B.8, for those $e_t > \alpha\varepsilon$, we should have:

$$t \leq \sum_{k=1}^K \mathbb{I}[e_k \geq e_t] \leq (\frac{\beta}{e_t^2} + 1)\text{dimE}_\alpha(\mathcal{F}, \varepsilon).$$

which implies $e_t \leq \sqrt{\frac{\beta\text{dimE}_\alpha(\mathcal{F},\varepsilon)}{t-\text{dimE}_\alpha(\mathcal{F},\varepsilon)}}$. Therefore, for any ε, we have:

$$\begin{aligned}
\sum_{k=1}^K e_k \leq& \alpha K\varepsilon + \sum_{k=1}^K \mathbb{I}[e_k > \alpha\varepsilon]e_k \\
\leq& \alpha K\varepsilon + (\text{dimE}_\alpha(\mathcal{F}, \varepsilon) + 1)C + \sum_{k=\text{dimE}_\alpha(\mathcal{F},\varepsilon)+2}^K \sqrt{\frac{\beta\text{dimE}_\alpha(\mathcal{F}, \varepsilon)}{t - \text{dimE}_\alpha(\mathcal{F}, \varepsilon)}} \\
& \text{(Recall the constant } C \text{ is the upper bound for } \mathbf{D}(f, f^*)(x)) \\
\leq& \alpha K\varepsilon + (\text{dimE}_\alpha(\mathcal{F}, \varepsilon) + 1)C + \sqrt{\beta\text{dimE}_\alpha(\mathcal{F}, \varepsilon)} \int\!\!\!\!\!\!\sum_{t=\text{dimE}_\alpha(\mathcal{F},\varepsilon)+1}^{K} \frac{1}{\sqrt{t - \text{dimE}_\alpha(\mathcal{F}, \varepsilon)}}dt \\
=& O(\sqrt{\beta K\text{dimE}_\alpha(\mathcal{F}, \varepsilon)} + \alpha K\varepsilon).
\end{aligned}$$

□

C Proofs for MLE Arguments

In this section, we only provide the proof for the MLE arguments of the algorithm flow for Mean Field Game, where in each iteration, we collect two data w.r.t. two policies in two modes. One can easily obtain the proof for the DCP of MFC by directly assigning $\widetilde{\pi} = \pi$ and removing the discussion for data $\{\widetilde{s}, \widetilde{a}, \widetilde{s}'\}$, so we omit it.

In the following, given the data collected at iteration k, $\mathcal{Z}^k := \{\{s_h^k, a_h^k, s_{h+1}'^k\}_{h=1}^H \cup \{\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{s}_{h+1}'^k\}_{h=1}^H\}$, we use $f_M^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)$ to denote the conditional probability w.r.t. model M with transition function $\{\mathbb{P}_{T,h}\}_{h=1}^H$, i.e.:

$$f_M^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k) = \prod_{h\in[H]} \mathbb{P}_{T,h}(s_{h+1}'^k | s_h^k, a_h^k, \mu_{M,h}^{\pi^k})\mathbb{P}_{T,h}(\widetilde{s}_{h+1}'^k | \widetilde{s}_h^k, \widetilde{a}_h^k, \mu_{M,h}^{\pi^k}).$$

For the simplicity of notations, we divide the random variables in $\mathcal{Z}^k$ into two parts depending on whether they are conditioned or not:

$$\mathcal{Z}_{cond}^k := \{(s_h^k, a_h^k)_{h=1}^H \cup (\widetilde{s}_h^k, \widetilde{a}_h^k)_{h=1}^H\}, \quad \mathcal{Z}_{pred}^k := \{(s_{h+1}'^k)_{h=1}^H \cup (\widetilde{s}_{h+1}'^k)_{h=1}^H\}.$$

Note that for different $h \in [H]$, $(s_h^k, a_h^k, s_{h+1}'^k)$ or $(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{s}_{h+1}'^k)$ are sampled from different trajectories. Therefore, there is no correlation between s_h^k, a_h^k (or $\widetilde{s}_h^k, \widetilde{a}_h^k$) with $s_{h'}'^k, a_{h'}'^k$ (or $\widetilde{s}_{h'}'^k, \widetilde{a}_{h'}'^k$) for those $h \neq h'$.

Lemma C.1. *In the following, for the data $\mathcal{Z}^1, ..., \mathcal{Z}^k$ collected in Alg. 1 in M^*, for any $\delta \in (0, 1)$:*

$$\Pr(\max_{M\in\mathcal{M}} \sum_{i=1}^k \log \frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)} \geq \log \frac{|\mathcal{M}|K}{\delta}) \leq \delta, \quad \forall k \in [K].$$

Proof. We denote $\mathbb{E}_k := \mathbb{E}[\cdot|\{(\pi^i, \widetilde{\pi}^i, \mathcal{Z}^i)\}_{i=1}^{k-1} \cup \{\pi^k, \widetilde{\pi}^k\}, M^*]$. First of all, for any $M \in \mathcal{M}$, we have:

$$\begin{aligned}
\mathbb{E}[\exp(\sum_{i=1}^k \log \frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)})] =& \mathbb{E}[\exp(\sum_{i=1}^{k-1} \log \frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)})\mathbb{E}_k[\exp(\log \frac{f_M^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}{f_{M^*}^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)})]] \\
=& \mathbb{E}[\exp(\sum_{i=1}^{k-1} \log \frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)})\mathbb{E}_k[\frac{f_M^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}{f_{M^*}^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}]]
\end{aligned}$$

$$
\begin{aligned}
&=\mathbb{E}[\exp(\sum_{i=1}^{k-1}\log\frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)})]\\
&=1.
\end{aligned}
$$

Here the last but two step is because:

$$
\begin{aligned}
\mathbb{E}_k[\frac{f_M^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}{f_{M^*}^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}] =&\mathbb{E}_{\mathcal{Z}^k_{cond}}[\mathbb{E}_{\mathcal{Z}^k_{pred}}[\frac{f_M^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}{f_{M^*}^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}|\mathcal{Z}^k_{cond},\boldsymbol{\mu}^{\pi^k}_{M^*},M^*]|\pi^k,\widetilde{\pi}^k,M^*]\\
=&\mathbb{E}_{\mathcal{Z}^k_{cond}}[\sum_{\mathcal{Z}^k_{pred}} f_{M^*}^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)\frac{f_M^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}{f_{M^*}^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)}|\pi^k,\widetilde{\pi}^k,M^*]\\
=&\mathbb{E}_{\mathcal{Z}^k_{cond}}[\sum_{\mathcal{Z}_{pred}} f_{M}^{\pi^k,\widetilde{\pi}^k}(\mathcal{Z}^k)|\pi^k,\widetilde{\pi}^k,M^*]=\mathbb{E}_{\mathcal{Z}^k_{cond}}[1||\pi^k,\widetilde{\pi}^k,M^*]=1.
\end{aligned}
$$

where $\sum_{\mathcal{Z}^k_{pred}}$ means summation over all possible value of $\mathcal{Z}^k_{pred}$.

Therefore, by Markov Inequality, for any fixed $M\in\mathcal{M}$ and fixed $k\in[K]$, and arbitrary $\delta\in(0,1)$, we have:

$$
\Pr(\sum_{i=1}^k\log\frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}\geq\log\frac{1}{\delta})\leq\delta\cdot\mathbb{E}[\exp(\sum_{i=1}^k\log\frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)})]=\delta.
$$

By taking union bound over all $M\in\mathcal{M}$ and all $k\in[K]$, we have:

$$
\Pr(\max_{M\in\mathcal{M}}\sum_{i=1}^k\log\frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}\geq\log\frac{|\mathcal{M}|K}{\delta})\leq\delta,\quad\forall k\in[K].
$$

□

Given dataset $D^k:=\{(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i)\}_{i=1}^k$, we use $\bar{D}^k$ to denote the "tangent" sequence $\{(\pi^i,\widetilde{\pi}^i,\bar{\mathcal{Z}}^i)\}_{i=1}^k$ where the policies are the same as D^k while each $\bar{\mathcal{Z}}^i$ is independently sampled from the same distribution as $\mathcal{Z}^i$ conditioning on π^i and $\widetilde{\pi}^i$.

Lemma C.2. *Let $l:\Pi\times\Pi\times(\mathcal{S}\times\mathcal{A}\times\mathcal{S})^H\times(\mathcal{S}\times\mathcal{A}\times\mathcal{S})^H\rightarrow\mathbb{R}$ be a real-valued loss function which maps from the joint space of two policies and space of $\mathcal{Z}^k$ to $\mathbb{R}$. Define $L(D^k):=\sum_{i=1}^k l(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i)$ and $L(\bar{D}^k):=\sum_{i=1}^k l(\pi^i,\widetilde{\pi}^i,\bar{\mathcal{Z}}^i)$. Then, for arbitrary $k\in[K]$,*

$$
\mathbb{E}[\exp(L(D^k)-\log\mathbb{E}[\exp(L(\bar{D}^k))|D^k])]=1.
$$

Proof. We denote $E^i:=\mathbb{E}_{\mathcal{Z}^i}[\exp(l(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i))|\pi^i,\widetilde{\pi}^i,M^*]$. By definition of $\bar{\mathcal{Z}}^i$, we should also have:

$$
\mathbb{E}_{\bar{D}^k}[\exp(\sum_{i=1}^k l(\pi^i,\widetilde{\pi}^i,\bar{\mathcal{Z}}^i))|D^k]=\prod_{i=1}^k E^i.
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{D^k}[\exp(L(D^k)-\log\mathbb{E}_{\bar{D}^k}[\exp(L(\bar{D}^k))|D^k])]\\
=&\mathbb{E}_{D^{k-1}\cup\{\pi^k,\widetilde{\pi}^k\}}[\mathbb{E}_{\mathcal{Z}^k}[\frac{\exp(\sum_{i=1}^k l(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i))}{\mathbb{E}_{\bar{D}^k}[\exp(\sum_{i=1}^k l(\pi^i,\widetilde{\pi}^i,\bar{\mathcal{Z}}^i))|D^k]}|D^{k-1}\cup\{\pi^k,\widetilde{\pi}^k\}]]\\
=&\mathbb{E}_{D^{k-1}\cup\{\pi^k,\widetilde{\pi}^k\}}[\mathbb{E}_{\mathcal{Z}^k}[\frac{\exp(\sum_{i=1}^k l(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i))}{\prod_{i=1}^k E^i}|D^{k-1}\cup\{\pi^k,\widetilde{\pi}^k\}]]\\
=&\mathbb{E}_{D^{k-1}\cup\{\pi^k,\widetilde{\pi}^k\}}[\frac{\exp(\sum_{i=1}^{k-1} l(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i))}{\prod_{i=1}^{k-1} E^i}\cdot\mathbb{E}_{\mathcal{Z}^k}[\frac{l(\pi^k,\widetilde{\pi}^k,\mathcal{Z}^k)}{E^k}|D^{k-1}\cup\{\pi^k,\widetilde{\pi}^k\}]]
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}_{D^{k-1}\cup\{\pi^k,\widetilde{\pi}^k\}}[\frac{\exp(\sum_{i=1}^{k-1} l(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i))}{\prod_{i=1}^{k-1} E^i}] \\
=&\mathbb{E}_{D^{k-1}}[\frac{\exp(\sum_{i=1}^{k-1} l(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i))}{\prod_{i=1}^{k-1} E^i}] = ... = 1.
\end{aligned}
$$

□

Theorem 4.2. *[Guarantees for MLE] By running Alg. 1 with any* $\delta \in (0,1)$*, with probability* $1-\delta$*, for all* $k \in [K]$*, we have* $M^* \in \widehat{\mathcal{M}}^k$*; for each* $M \in \widehat{\mathcal{M}}^k$ *with transition* $\mathbb{P}_T$ *and any* $h \in [H]$*:*

$$
\sum_{i=1}^k \mathbb{E}_{\pi^i,M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s_h^i,a_h^i,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|s_h^i,a_h^i,\mu^{\pi^i}_{M^*,h}))] \leq 2\log(\frac{2|\mathcal{M}|KH}{\delta}).
$$

Besides, for MFG branch, we additionally have:

$$
\sum_{i=1}^k \mathbb{E}_{\widetilde{\pi}^i,M^*|\boldsymbol{\mu}^{\pi^i}_{M^*}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\widetilde{s}_h^i,\widetilde{a}_h^i,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|\widetilde{s}_h^i,\widetilde{a}_h^i,\mu^{\pi^i}_{M^*,h}))] \leq 2\log(\frac{2|\mathcal{M}|KH}{\delta}).
$$

Proof. Given a model $M \in \mathcal{M}$, we consider the loss function:

$$
l_M(\pi,\widetilde{\pi},\mathcal{Z}) := \begin{cases} \log\frac{f_M^{\pi,\widetilde{\pi}}(\mathcal{Z})}{f_{M^*}^{\pi,\widetilde{\pi}}(\mathcal{Z})}, & \text{if } f_{M^*}^{\pi,\widetilde{\pi}}(\mathcal{Z}) \neq 0 \\ 0, & \text{otherwise} \end{cases}
$$

Define $M^k_{\text{MLE}} \leftarrow \arg\max_{M\in\mathcal{M}} l^k_{\text{MLE}}(M)$. Considering the event $\mathcal{E}$:

$$
\mathcal{E} := \{l^k_{\text{MLE}}(M^k_{\text{MLE}}) - l^k_{\text{MLE}}(M^*) \leq \log\frac{2|\mathcal{M}|KH}{\delta}, \quad \forall k \in [K]\}.
$$

and the event $\mathcal{E}'$ defined by:

$$
\mathcal{E}' := \{-\log\mathbb{E}_{\bar{D}^k}[\exp L_M(\bar{D}^k)|D^k] \leq -L_M(D^k) + \log(\frac{2|\mathcal{M}|KH}{\delta}), \quad \forall M \in \mathcal{M}, k \in [K]\}.
$$

where we define $L_M(D^k) := \sum_{i=1}^k l_M(\pi^i,\widetilde{\pi}^i,\mathcal{Z}^i)$ and $L_M(\bar{D}^k) := \sum_{i=1}^k l_M(\pi^i,\widetilde{\pi}^i,\bar{\mathcal{Z}}^i)$. By Lem. C.1, we have $\Pr(\mathcal{E}) \geq 1-\frac{\delta}{2H}$. Besides, by applying Lem. C.2 on l_M defined above and applying Markov inequality and the union bound over all $M \in \mathcal{M}$ and $k \in [K]$, we have $\Pr(\mathcal{E}') \geq 1-\frac{\delta}{2H}$.

On the event $\mathcal{E}\cap\mathcal{E}'$, for any $k \in [K]$, we have $M^* \in \widehat{\mathcal{M}}^k$, and for any $M \in \widehat{\mathcal{M}}^k$:

$$
\begin{aligned}
-\log\mathbb{E}_{\bar{D}^k}[\exp L_M(\bar{D}^k)|D^k] \leq & -L_M(D^k) + \log(\frac{2|\mathcal{M}|KH}{\delta}) \\
=& l^k_{\text{MLE}}(M^*) - l^k_{\text{MLE}}(M) + \log(\frac{2|\mathcal{M}|KH}{\delta}) \\
\leq & l^k_{\text{MLE}}(M^k_{\text{MLE}}) - l^k_{\text{MLE}}(M) + \log(\frac{2|\mathcal{M}|KH}{\delta}) \\
\leq & 2\log(\frac{2|\mathcal{M}|KH}{\delta}).
\end{aligned}
$$

Therefore, for any k and any $M \in \widehat{\mathcal{M}}^k$,

$$
\begin{aligned}
2\log(\frac{2|\mathcal{M}|KH}{\delta}) \geq & -\sum_{i=1}^k \log\mathbb{E}_{\mathcal{Z}^i}[\sqrt{\frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}}|\pi^i,\widetilde{\pi}^i,M^*] \\
\geq & \sum_{i=1}^k 1 - \mathbb{E}_{\mathcal{Z}^i}[\sqrt{\frac{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}{f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}}|\pi^i,\widetilde{\pi}^i,M^*] \qquad (-\log x \geq 1-x)
\end{aligned}
$$

$$=\sum_{i=1}^{k}\mathbb{E}_{\mathcal{Z}^i_{cond}}[1-\sum_{\mathcal{Z}^i_{pred}}\sqrt{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}|\pi^i,\widetilde{\pi}^i,M^*].$$

For any $i\in[k]$ and for arbitrary random variable $s^i_h, a^i_h \in \mathcal{Z}^i_{cond}$ and $s'^i_{h+1}\in\mathcal{Z}^i_{pred}$, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{Z}^i_{cond}}[1-\sum_{\mathcal{Z}^i_{pred}}\sqrt{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}|\pi^i,\widetilde{\pi}^i,M^*]\\
=&\mathbb{E}_{\mathcal{Z}^i_{cond}}[1-\sum\!\!\!\!\!\!\int_{s'^i_{h+1}}\sqrt{\mathbb{P}_{T,h}(s'^i_{h+1}|s^i_h,a^i_h,\mu^{\pi^i}_{M,h})\mathbb{P}_{T^*,h}(s'^i_{h+1}|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h})}\sum_{\mathcal{Z}^i_{pred}\setminus\{s'^i_{h+1}\}}\sqrt{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}|\pi^i,\widetilde{\pi}^i,M^*]\\
&\qquad\qquad\text{(Independence between } s'^i_{h+1} \text{ and } \mathcal{Z}^i\setminus\{s'^i_{h+1}\} \text{ conditioning on } \mathcal{Z}^i_{cond})\\
\geq&\mathbb{E}_{\mathcal{Z}^i_{cond}}[1-\sum\!\!\!\!\!\!\int_{s'^i_{h+1}}\sqrt{\mathbb{P}_{T,h}(s'^i_{h+1}|s^i_h,a^i_h,\mu^{\pi^i}_{M,h})\mathbb{P}_{T^*,h}(s'^i_{h+1}|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h})}|\pi^i,\widetilde{\pi}^i,M^*]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad(\sqrt{ab}\leq\tfrac{a+b}{2})\\
=&\mathbb{E}_{s^i_h,a^i_h}[1-\sum\!\!\!\!\!\!\int_{s'^i_{h+1}}\sqrt{\mathbb{P}_{T,h}(s'^i_{h+1}|s^i_h,a^i_h,\mu^{\pi^i}_{M,h})\mathbb{P}_{T^*,h}(s'^i_{h+1}|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h})}|\pi^i,\widetilde{\pi}^i,M^*]\\
=&\mathbb{E}_{\pi^i,M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}))].
\end{aligned}$$

Similarly, for arbitrary random variable $\widetilde{s}^i_h,\widetilde{a}^i_h\in\mathcal{Z}^i_{cond}$ and $\widetilde{s}'^i_{h+1}\in\mathcal{Z}^i_{pred}$, we have:

$$\mathbb{E}_{\mathcal{Z}^i_{cond}}[1-\sum_{\mathcal{Z}^i_{pred}}\sqrt{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}|\pi^i,\widetilde{\pi}^i,M^*]\geq\mathbb{E}_{\widetilde{\pi}^i,M^*|\boldsymbol{\mu}^{\pi^i}_{M^*}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M^*,h}))].$$

Therefore, on the event $\mathcal{E}'$, for any $k\in[K]$, $M\in\widehat{\mathcal{M}}^k$, and a fixed $h\in[H]$, we have:

$$\begin{aligned}
2\log(\frac{2|\mathcal{M}|KH}{\delta})\geq&\sum_{i=1}^{k}\mathbb{E}_{\pi^i,M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}))]\\
2\log(\frac{2|\mathcal{M}|KH}{\delta})\geq&\sum_{i=1}^{k}\mathbb{E}_{\widetilde{\pi}^i,M^*|\boldsymbol{\mu}^{\pi^i}_{M^*}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M^*,h}))].
\end{aligned}$$

By taking the union bound for all $h\in[H]$, we finish the proof for DCP of MFG. The analysis and results for MFC is similar and easier so we omit it here. □

Corollary C.3. *Under the same event in Thm. 4.2, for any $k\in[K]$, $M\in\widehat{\mathcal{M}}^k$, and a fixed $h\in[H]$, we have:*

$$\begin{aligned}
\sum_{i=1}^{k}\mathbb{E}_{\pi^i,M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}))]\leq&(4+8L_T^2H^2)\log(\frac{2|\mathcal{M}|KH}{\delta}),\\
\sum_{i=1}^{k}\mathbb{E}_{\widetilde{\pi}^i,M^*|\boldsymbol{\mu}^{\pi^i}_{M^*}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M^*,h}),\ \mathbb{P}_{T^*,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M^*,h}))]\leq&(4+8L_T^2H^2)\log(\frac{2|\mathcal{M}|KH}{\delta}).
\end{aligned}$$

Proof. By Assump. B, for any i, we have:

$$\begin{aligned}
&\mathbb{E}_{\pi^i,M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}))]\\
\leq&2\mathbb{E}_{\pi^i,M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}))]+2L_T^2\|\mu^{\pi^i}_{M,h}-\mu^{\pi^i}_{M^*,h}\|^2_{\mathbb{TV}}\\
\leq&2\mathbb{E}_{\pi^i,M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h,a^i_h,\mu^{\pi^i}_{M^*,h}))]\\
&\quad+4L_T^2H\mathbb{E}_{\pi,M}[\sum_{h'=1}^{h-1}\mathbb{H}^2(\mathbb{P}_{T,h'}(\cdot|s^i_{h'},a^i_{h'},\mu^{\pi^i}_{M,h'}),\ \mathbb{P}_{T^*,h'}(\cdot|s^i_{h'},a^i_{h'},\mu^{\pi^i}_{M^*,h'}))].\\
&\qquad\qquad\qquad\qquad\text{(Lem. D.1; Cauchy-Schwarz inequality; } \|P-Q\|_{\mathbb{TV}}\leq\sqrt{2}\mathbb{H}(P,Q))
\end{aligned}$$

Therefore, on the event $\mathcal{E}'$, for any $k \in [K]$, $M \in \widehat{\mathcal{M}}^k$, and a fixed $h \in [H]$, we have:

$$\sum_{i=1}^{k} \mathbb{E}_{\pi^i, M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s_h^i, a_h^i, \mu_{M^*,h}^{\pi^i}),\ \mathbb{P}_{T^*,h}(\cdot|s_h^i, a_h^i, \mu_{M^*,h}^{\pi^i}))] \leq (4 + 8L_T^2 H^2) \log(\frac{2|\mathcal{M}|KH}{\delta}).$$

Similarly, we have:

$$\begin{aligned}&\mathbb{E}_{\widetilde{\pi}^i, M^*|\boldsymbol{\mu}_{M^*}^{\pi^i}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M,h}^{\pi^i}),\ \mathbb{P}_{T^*,h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M^*,h}^{\pi^i}))] \\ \leq& 2\mathbb{E}_{\widetilde{\pi}^i, M^*|\boldsymbol{\mu}_{M^*}^{\pi^i}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M^*,h}^{\pi^i}),\ \mathbb{P}_{T^*,h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M^*,h}^{\pi^i}))] + 2L_T^2 \|\mu_{M,h}^{\pi^i} - \mu_{M^*,h}^{\pi^i}\|_{\mathbb{TV}}^2.\end{aligned}$$

By similar discussion, we have:

$$\sum_{i=1}^{k} \mathbb{E}_{\widetilde{\pi}^i, M^*|\boldsymbol{\mu}_{M^*}^{\pi^i}}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M^*,h}^{\pi^i}),\ \mathbb{P}_{T^*,h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M^*,h}^{\pi^i}))] \leq (4 + 8L_T^2 H^2) \log(\frac{2|\mathcal{M}|KH}{\delta}).$$

□

Theorem C.4. *[Accumulative Model Difference] For any $\delta \in (0,1)$, with probability $1 - 3\delta$, for any sequence $\{\widehat{M}^{k+1}\}_{k\in[K]}$ with $\widehat{M}^{k+1} \in \widehat{\mathcal{M}}^{k+1}$ for all $k \in [K]$, and any $h \in [H]$, we have:*

$$\begin{aligned}&\sum_{k=1}^{K} \mathbb{E}_{\pi^{k+1}, M^*}[\|\mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi^{k+1}}) - \mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi^{k+1}})\|_{\mathbb{TV}}] \\ &\qquad = O\Big((1 + L_T H)\sqrt{K dimE_\alpha(\mathcal{M}, \varepsilon_0) \log \frac{2|\mathcal{M}|KH}{\delta}} + \alpha K \varepsilon_0\Big) \\ &\sum_{k=1}^{K} \mathbb{E}_{\widetilde{\pi}^{k+1}, M^*|\boldsymbol{\mu}_{M^*}^{\pi^{k+1}}}[\|\mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi^{k+1}}) - \mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi^{k+1}})\|_{\mathbb{TV}}] \\ &\qquad = O\Big((1 + L_T H)\sqrt{K dimE_\alpha(\mathcal{M}, \varepsilon_0) \log \frac{2|\mathcal{M}|KH}{\delta}} + \alpha K \varepsilon_0\Big).\end{aligned}$$

Proof. We first take a look at the data $(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{s}_{h+1}'^k)$ collected by $(\widetilde{\pi}^i, \pi^i)$ and the Eluder Dimension w.r.t. the Hellinger distance. On the event in Thm. 4.2 and Lem. D.4, there exists an absolute constant $c_{\mathbb{H}}$, s.t., w.p. $1 - \frac{\delta}{2}$, for any $h \in [H]$, and any $\widehat{M}^{k+1} \in \widehat{\mathcal{M}}^{k+1}$, we have:

$$\sum_{i=1}^{k} \mathbb{H}^2(\mathbb{P}_{T^*,h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M^*,h}^{\pi^i}), \mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|\widetilde{s}_h^i, \widetilde{a}_h^i, \mu_{M^*,h}^{\pi^i})) \leq c_{\mathbb{H}}(1 + L_T^2 H^2) \log \frac{2|\mathcal{M}|KH}{\delta}. \tag{10}$$

By Lem. 4.4, there exists some constant $c'_{\mathbb{H}}$, for any ε_0, we have:

$$\begin{aligned}&\sum_{k=1}^{K} \mathbb{H}(\mathbb{P}_{T^*,h}(\cdot|\widetilde{s}_h^{k+1}, \widetilde{a}_h^{k+1}, \mu_{M^*,h}^{\pi^{k+1}}), \mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|\widetilde{s}_h^{k+1}, \widetilde{a}_h^{k+1}, \mu_{M^*,h}^{\pi^{k+1}})) \\ &\qquad\qquad \leq c'_{\mathbb{H}}\Big((1 + L_T H)\sqrt{K \mathrm{dimE}_\alpha(\mathcal{M}, \mathbb{H}, \varepsilon_0) \log \frac{2|\mathcal{M}|KH}{\delta}} + \alpha K \varepsilon_0\Big).\end{aligned}$$

By applying Lem. D.4 again, w.p. $1 - \frac{\delta}{2}$, we have:

$$\begin{aligned}&\sum_{k=1}^{K} \mathbb{E}_{\widetilde{\pi}^{k+1}, M^*|\boldsymbol{\mu}_{M^*}^{\pi^{k+1}}}[\mathbb{H}(\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi^{k+1}}), \mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|s_h, a_h, \mu_{M^*,h}^{\pi^{k+1}}))] \\ \leq& 3c'_{\mathbb{H}}\Big((1 + L_T H)\sqrt{K \mathrm{dimE}_\alpha(\mathcal{M}, \mathbb{H}, \varepsilon_0) \log \frac{2|\mathcal{M}|KH}{\delta}} + \alpha K \varepsilon_0\Big) + \log \frac{2|\mathcal{M}|H}{\delta} \\ \leq& (3c'_{\mathbb{H}} + 1)\Big((1 + L_T H)\sqrt{K \mathrm{dimE}_\alpha(\mathcal{M}, \mathbb{H}, \varepsilon_0) \log \frac{2|\mathcal{M}|KH}{\delta}} + \alpha K \varepsilon_0\Big).\end{aligned}$$

By the relationship that $\mathbb{TV}(P,Q) \leq \sqrt{2}\mathbb{H}(P,Q)$, on the one hand, the above implies an upper bound for the accumulative model difference measured by TV-distance; on the other hand, we can conduct similar discussion for the Eluder dimension $\text{dimE}(\mathcal{M},\mathbb{TV},\varepsilon_0)$ and derive another upper bound. Combine them together, for some constant c, we have:

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}_{\widetilde{\pi}^{k+1},M^*|\boldsymbol{\mu}^{\pi^{k+1}}_{M^*}}[\|\mathbb{P}_{T^*,h}(\cdot|s_h,a_h,\mu^{\pi^{k+1}}_{M^*,h}),\mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|s_h,a_h,\mu^{\pi^{k+1}}_{M^*,h})\|_{\mathbb{TV}}]\\ \leq&(3c+1)\Big((1+L_TH)\sqrt{K\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}}+\alpha K\varepsilon_0\Big).\end{aligned}$$

where we use that $\dim_\alpha(\mathcal{M},\varepsilon_0)=\min\{\text{dimE}_\alpha(\mathcal{M},\mathbb{H},\varepsilon_0),\text{dimE}_\alpha(\mathcal{M},\mathbb{TV},\varepsilon_0)\}$.

Then, we can conduct similar discussion for the data (s_h^i,a_h^i,s_{h+1}^i) collected by (π^{k+1},π^{k+1}), and for some constant c', we have:

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}_{\pi^{k+1},M^*}[\|\mathbb{P}_{T^*,h}(\cdot|s_h,a_h,\mu^{\pi^{k+1}}_{M^*,h}),\mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|s_h,a_h,\mu^{\pi^{k+1}}_{M^*,h})\|_{\mathbb{TV}}]\\ \leq&(3c'+1)\Big((1+L_TH)\sqrt{K\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}}+\alpha K\varepsilon_0\Big).\end{aligned}$$

We finish the proof by noting that the total failure rate can be upper bounded by $\delta+\delta/2\cdot 2\cdot 2=3\delta$. □

D Proofs for Mean-Field Reinforcement Learning

D.1 Missing Details

Assumption D (Contraction Operator)**.** For arbitrary h, and arbitrary valid density $\mu_h,\mu'_h\in\Delta(\mathcal{S})$, and arbitrary model $M:=(\mathbb{P}_T,\mathbb{P}_r)\in\mathcal{M}$, there exists $L_\Gamma<1$, such that,

$$\|\Gamma^\pi_{M,h}(\mu_h)-\Gamma^\pi_{M,h}(\mu'_h)\|_{\mathbb{TV}}\leq L_\Gamma\|\mu_h-\mu'_h\|_{\mathbb{TV}}.$$

where $\Gamma^\pi_{M,h}(\mu_h)$ is defined in Eq. (1). According to [63], Assump. D is implied by some Lipschitz continuous assumption on the transition function w.r.t. the Dirac distance $d(s,s'):=\mathbb{I}[s\neq s']$ (at least when $\mathcal{S}$ and $\mathcal{A}$ are countable). As we will see later, although Assump. D is not necessary to derive sample complexity bound, it can be useful to get rid of exponential dependence on L_T.

Algorithm 3: Regret to PAC Conversion

```
Input: Policy sequence $\pi^1,...,\pi^K$; Accuracy level $\varepsilon$; Confidence level $\delta$.
$N \leftarrow \lceil\log_{\frac{3}{2}}\frac{1}{\delta}\rceil$.
Randomly select $N$ policies from $\pi^1,...,\pi^K$, denoted as $\pi^{k_1},...\pi^{k_N}$.
for $n\in[N]$ do
    Sample $\frac{16}{\varepsilon^2}\log\frac{2N}{\delta}$ trajectories by deploying $\pi^{k_n}$.
    Compute empirical estimation $\widehat{J}_{M^*}(\pi^{k_n})$ by averaging the return in trajectories.
end
return $\pi:=\pi^{k_{n^*}}$ with $n^*\leftarrow\arg\max_{n\in[N]}\widehat{J}_{M^*}(\pi^{k_n})$.
```

D.2 Proofs for Basic Lemma

Lemma D.1. *[Density Estimation Error] Given two model M and $\widetilde{M}$ and a policy π, we have:*

$$\|\mu^\pi_{M,h+1}-\mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}}\leq\mathbb{E}_{\pi,M}[\sum_{h'=1}^{h}\|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})-\mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{\widetilde{M},h'})\|_{\mathbb{TV}}]. \tag{11}$$

Besides, under Assump. B, we have:

$$\|\mu^\pi_{M,h+1} - \mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}} \le \mathbb{E}_{\pi,M}[\sum_{h'=1}^{h} (1+L_T)^{h-h'} \|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'}) - \mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})\|_{\mathbb{TV}}]. \tag{12}$$

Moreover, under Assump. B and Assump. D, we have:

$$\|\mu^\pi_{M,h+1} - \mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}} \le \mathbb{E}_{\pi,M}[\sum_{h'=1}^{h} L_\Gamma^{h-h'} \|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'}) - \mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})\|_{\mathbb{TV}}]. \tag{13}$$

Proof. In the following, we will use $\bar{\mathcal{S}}$ or $\bar{\mathcal{S}}'$ to denote a subset of $\mathcal{S}$.

Proof for Eq. (11)

$$\begin{aligned}
&\|\mu^\pi_{M,h+1} - \mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}} \\
=&\sup_{\bar{\mathcal{S}}\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_{h+1}\in\bar{\mathcal{S}}} \Big(\sum\!\!\!\!\!\!\int_{s_h,a_h} \mu^\pi_{M,h}(s_h)\pi(a_h|s_h)\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h}) - \sum\!\!\!\!\!\!\int_{s_h,a_h} \mu^\pi_{\widetilde{M},h}(s_h)\pi(a_h|s_h)\mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{\widetilde{M},h})\Big)| \\
=&\sup_{\bar{\mathcal{S}}\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_{h+1}\in\bar{\mathcal{S}}} \sum\!\!\!\!\!\!\int_{s_h,a_h} (\mu^\pi_{M,h}(s_h) - \mu^\pi_{\widetilde{M},h}(s_h))\pi(a_h|s_h)\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h})| \\
&+ \sup_{\bar{\mathcal{S}}'\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_{h+1}\in\bar{\mathcal{S}}'} \sum\!\!\!\!\!\!\int_{s_h,a_h} \mu^\pi_{\widetilde{M},h}(s_h)\pi(a_h|s_h)(\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{\widetilde{M},h}))|.
\end{aligned}$$

For the first part, we have:

$$\begin{aligned}
&\sup_{\bar{\mathcal{S}}\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_{h+1}\in\bar{\mathcal{S}}} \sum\!\!\!\!\!\!\int_{s_h,a_h} (\mu^\pi_{M,h}(s_h) - \mu^\pi_{\widetilde{M},h}(s_h))\pi(a_h|s_h)\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h})| \\
\le&\sup_{\bar{\mathcal{S}}\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_h}(\mu^\pi_{M,h}(s_h) - \mu^\pi_{\widetilde{M},h}(s_h)) \sum\!\!\!\!\!\!\int_{a_h}\pi(a_h|s_h) \sum\!\!\!\!\!\!\int_{s_{h+1}\in\bar{\mathcal{S}}} \mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h})| \\
\le&\sup_{\bar{\mathcal{S}}\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_h\in\bar{\mathcal{S}}} \mu^\pi_{M,h}(s_h) - \mu^\pi_{\widetilde{M},h}(s_h)| \\
=&\|\mu^\pi_{M,h} - \mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}}.
\end{aligned}$$

For the second part, we have:

$$\begin{aligned}
&\sup_{\bar{\mathcal{S}}'\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_{h+1}\in\bar{\mathcal{S}}'} \sum\!\!\!\!\!\!\int_{s_h,a_h} \mu^\pi_{\widetilde{M},h}(s_h)\pi(a_h|s_h)(\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{\widetilde{M},h}))| \\
\le&\sum\!\!\!\!\!\!\int_{s_h,a_h} \mu^\pi_{\widetilde{M},h}(s_h)\pi(a_h|s_h) \sup_{\bar{\mathcal{S}}'\subset\mathcal{S}} | \sum\!\!\!\!\!\!\int_{s_{h+1}\in\bar{\mathcal{S}}'} (\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{\widetilde{M},h}))| \\
=&\mathbb{E}_{s_h\sim\mu^\pi_{M,h},a_h\sim\pi(\cdot|s_h)}[\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mu^\pi_{M,h+1} - \mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}} \le&\|\mu^\pi_{M,h} - \mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}} + \mathbb{E}_{s_h\sim\mu^\pi_{M,h},a_h\sim\pi(\cdot|s_h)}[\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}] \\
\le&... \le \mathbb{E}_{\pi,M}[\sum_{h'=1}^{h} \|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'}) - \mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{\widetilde{M},h'})\|_{\mathbb{TV}}].
\end{aligned} \tag{14}$$

Proof for Eq. (12) Starting with the first inequality of Eq. (14) and applying the Assump. B, we directly have:

$$\|\mu^\pi_{M,h+1} - \mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}} \le(1+L_T)\|\mu^\pi_{M,h} - \mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}} + \mathbb{E}_{s_h\sim\mu^\pi_h,a_h\sim\pi}[\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]$$

$$\leq \mathbb{E}_{\pi}[\sum_{h'=1}^{h}(1+L_T)^{h-h'}\|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})-\mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})\|_{\mathbb{TV}}].$$

Proof for Eq. (13) Under Assump. D, we can use a different way to decompose the density difference.

$$\begin{aligned}
&\|\mu^\pi_{M,h+1}-\mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}}\\
=&\sup_{\bar{\mathcal{S}}\subset\mathcal{S}}|\mathop{\sum\!\!\!\!\!\!\int}_{s_{h+1}\in\bar{\mathcal{S}}}\Big(\mathop{\sum\!\!\!\!\!\!\int}_{s_h,a_h}\mu^\pi_{M,h}(s_h)\pi(a_h|s_h)\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h})-\mathop{\sum\!\!\!\!\!\!\int}_{s_h,a_h}\mu^\pi_{\widetilde{M},h}(s_h)\pi(a_h|s_h)\mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{\widetilde{M},h})\Big)|\\
=&\sup_{\bar{\mathcal{S}}\subset\mathcal{S}}|\mathop{\sum\!\!\!\!\!\!\int}_{s_{h+1}\in\bar{\mathcal{S}}}\Big(\mathop{\sum\!\!\!\!\!\!\int}_{s_h,a_h}\mu^\pi_{M,h}(s_h)\pi(a_h|s_h)\mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h})-\mathop{\sum\!\!\!\!\!\!\int}_{s_h,a_h}\mu^\pi_{\widetilde{M},h}(s_h)\pi(a_h|s_h)\mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{\widetilde{M},h})\Big)|\\
&+\sup_{\bar{\mathcal{S}}\subset\mathcal{S}}|\mathop{\sum\!\!\!\!\!\!\int}_{s_{h+1}\in\bar{\mathcal{S}}}\mathop{\sum\!\!\!\!\!\!\int}_{s_h,a_h}\mu^\pi_{M,h}(s_h)\pi(a_h|s_h)\Big(\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h})\Big)|\\
\leq&\|\Gamma^\pi_{\widetilde{M},h}(\mu^\pi_{M,h})-\Gamma^\pi_{\widetilde{M},h}(\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}+\mathbb{E}_{s_h\sim\mu^\pi_{M,h},a_h\sim\pi}[\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]\\
\leq&L_\Gamma\|\mu^\pi_{M,h}-\mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}}+\mathbb{E}_{s_h\sim\mu^\pi_h,a_h\sim\pi}[\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]\\
\leq&\mathbb{E}_\pi[\sum_{h'=1}^{h}L_\Gamma^{h-h'}\|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})-\mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})\|_{\mathbb{TV}}].
\end{aligned}$$

□

As implied by Lem. D.1, we have the following corollary.

Corollary D.2. *In general,*

$$\sum_{h=1}^{H}\|\mu^\pi_{M,h}-\mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}}\leq\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}(H-h)\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}].$$

Besides, under Assump. B, we have:

$$\begin{aligned}
\sum_{h=1}^{H}\|\mu^\pi_{M,h}-\mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}}\leq&\sum_{h=1}^{H}\mathbb{E}_{\pi,M}[\sum_{h'=1}^{h-1}(1+L_T)^{h-h'-1}\|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})-\mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})\|_{\mathbb{TV}}]\\
=&\sum_{h=1}^{H}\frac{(1+L_T)^{H-h}-1}{L_T}\mathbb{E}_{\pi,M}[\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]
\end{aligned}$$

Moreover, with additional Assump. D, we have:

$$\begin{aligned}
\sum_{h=1}^{H}\|\mu^\pi_{M,h}-\mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}}\leq&\sum_{h=1}^{H}\mathbb{E}_{\pi,M}[\sum_{h'=1}^{h-1}L_\Gamma^{h-h'-1}\|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})-\mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})\|_{\mathbb{TV}}]\\
\leq&\sum_{h=1}^{H}\frac{1}{1-L_\Gamma}\mathbb{E}_{\pi,M}[\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}].
\end{aligned}$$

$$(L_\Gamma<1)$$

Theorem D.3 (Transition Difference Transformation; Full Version of Thm. 4.3)**.** *Given two arbitrary model $M=(\mathcal{S},\mathcal{A},H,\mathbb{P}_T,\mathbb{P}_r)$ and $\widetilde{M}=(\mathcal{S},\mathcal{A},H,\mathbb{P}_{\widetilde{T}},\mathbb{P}_r)$, and arbitrary policy π, under Assump. B, we have:*

$$\begin{aligned}
&\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]\\
&\qquad\leq(1+L_TH)\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}],
\end{aligned}\tag{15}$$

and

$$\begin{aligned}&\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]\\&\qquad\leq\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}(1+L_T)^{H-h}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}].\end{aligned}\tag{16}$$

Moreover, additionally under Assump. D, we have:

$$\begin{aligned}&\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]\\&\qquad\leq(1+\frac{L_T}{1-L_\Gamma})\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}].\end{aligned}\tag{17}$$

Proof. By Assump. B, we have:

$$\begin{aligned}&\Big|\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}]\\&\quad-\mathbb{E}_{\pi,M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]\Big|\leq L_T\sum_{h=1}^{H}\|\mu^\pi_{M,h}-\mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}}.\end{aligned}\tag{18}$$

Then, by applying Corollary D.2, and plugging into the above equation, we can finish the proof. □

Lemma D.4 (Concentration Lemma)**.** *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:*

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

Proof. Define $Z_k := \mathbb{E}[\exp(t\sum_{i=1}^k X_i - 3Y_i)]$. By taking $t \in [0, 1/C]$, we have:

$$\begin{aligned}\mathbb{E}[Z_k|\mathcal{F}_k]=&\exp(t\sum_{i=1}^{k-1}(X_i-3Y_i))\mathbb{E}[\exp(t(X_k-3Y_k))|\mathcal{F}_k]\\\leq&\exp(t\sum_{i=1}^{k-1}(X_i-3Y_i))\exp(-3Y_k)\mathbb{E}[1+tX_k+2t^2X_k^2|\mathcal{F}_k]\\\leq&\exp(t\sum_{i=1}^{k-1}(X_i-3Y_i))\exp(-3Y_k)\cdot(1+3tY_k) && (0\geq tX_k\leq 1)\\\leq&\exp(t\sum_{i=1}^{k-1}(X_i-3Y_i))\cdot\exp(-3Y_k+3tY_k) && (1+x\leq\exp(x))\\\leq&\exp(t\sum_{i=1}^{k-1}(X_i-3Y_i))=Z_{k-1}.\end{aligned}$$

We augment the sequence by set $X_0 = Y_0 = 0$, which implies $Z_0 = 1$. Therefore, $\{Z_k\}_{k\geq 0}$ is a super-martingale w.r.t. $\{\mathcal{F}_k\}_{k\geq 1}$. Denote τ to be the smallest t such that $\sum_{i=1}^t(X_i - 3Y_i) > C\log\frac{1}{\delta}$, we have:

$$Z_{k\wedge\tau}=\mathbb{E}[\exp(t\sum_{i=1}^{k\wedge\tau}(X_i-3Y_i))]$$

$$
\begin{aligned}
&=\mathbb{E}[\sum_{j=1}^{k}\mathbb{I}[\tau=j]\exp(t\sum_{i=1}^{\tau}(X_i-3Y_i))]+\mathbb{E}[\mathbb{I}[\tau>k]\exp(t\sum_{i=1}^{k}(X_i-3Y_i))]\\
&\leq\exp(tC)\mathbb{E}[\sum_{j=1}^{k}\mathbb{I}[\tau=j]\exp(t\sum_{i=1}^{\tau-1}(X_i-3Y_i))]+\mathbb{E}[\mathbb{I}[\tau>k]\exp(t\sum_{i=1}^{k}(X_i-3Y_i))]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad(\exp(t(X_i-3Y_i))\leq\exp(tC))\\
&\leq\exp(tC+tC\log\frac{1}{\delta})\sum_{j=1}^{k}\mathbb{E}[\mathbb{I}[\tau=j]]+\exp(tC\log\frac{1}{\delta})\mathbb{E}[\mathbb{I}[\tau>k]]\\
&\leq\exp(tC+tC\log\frac{1}{\delta}).
\end{aligned}
$$

which is upper bounded. Therefore, by the optimal stopping theorem, and choosing $t=1/C$, we have:

$$
\begin{aligned}
\Pr(\exists k\leq K,\ \sum_{i=1}^{k}X_k-3Y_k\geq C\log\frac{1}{\delta})=&\Pr(\tau\leq K)\leq\Pr(Z_{K\wedge\tau}\geq\exp(tl\log\frac{1}{\delta}))\\
\leq&\frac{\mathbb{E}[Z_{K\wedge\tau}]}{\exp(tC\log\frac{1}{\delta})}\leq\frac{Z_0}{\exp(tC\log\frac{1}{\delta})}=\delta.
\end{aligned}
$$

Since the above holds for arbitrary K, by setting $K\rightarrow+\infty$, we have:

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

The other inequality can be proved similarly by considering $Z'_k=\mathbb{E}[\exp(t\sum_{i=1}^{k}(Y_k-3X_k)]$. □

D.3 Proofs for RL for Mean-Field Control

Lemma 4.5. *[Simulation Lemma for MFC] Given an arbitrary model M with transition function $\mathbb{P}_T$, and an arbitrary policy π, under Assump. B, we have:*

$$
|J_{M^*}(\pi)-J_M(\pi)|\leq\mathbb{E}_{\pi,M^*}[\sum_{h=1}^{H}(1+L_rH)\|\mathbb{P}_{T^*,h}(\cdot|s_h,a_h,\mu^\pi_{M^*,h})-\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})\|_{\mathbb{TV}}].
$$

Proof. We first prove the value difference for the general case. The lemma can be proved by directly assign $\widetilde{M}=M^*$ and $\pi=\widetilde{\pi}$.

$$
\begin{aligned}
&|J_M(\widetilde{\pi};\boldsymbol{\mu}^\pi_M)-J_{\widetilde{M}}(\widetilde{\pi};\boldsymbol{\mu}^\pi_{\widetilde{M}})|\\
=&|\mathbb{E}_{s_1\sim\mu_1}[V^{\widetilde{\pi}}_{M,1}(s_1;\boldsymbol{\mu}^\pi_M)-V^{\widetilde{\pi}}_{\widetilde{M},1}(s_1;\boldsymbol{\mu}^\pi_{\widetilde{M}})]|\\
=&|\mathbb{E}_{s_1\sim\mu_1,a_1\sim\widetilde{\pi}}[r_1(s_1,a_1,\mu^\pi_{M,1})-r_1(s_1,a_1,\mu^\pi_{\widetilde{M},1})\\
&\qquad+\sum\!\!\!\!\!\!\!\int_{s_2}\mathbb{P}_{T,1}(s_2|s_1,a_1,\mu^\pi_{M,1})V^{\widetilde{\pi}}_{M,2}(s_2;\boldsymbol{\mu}^\pi_M)-\sum\!\!\!\!\!\!\!\int_{s_2}\mathbb{P}_{\widetilde{T},1}(s_2|s_1,a_1,\mu^\pi_{\widetilde{M},1})V^{\widetilde{\pi}}_{\widetilde{M},2}(s_2;\boldsymbol{\mu}^\pi_{\widetilde{M}})]|\\
\leq&L_r\|\mu^\pi_{M,1}-\mu^\pi_{\widetilde{M},1}\|_{\mathbb{TV}}+|\mathbb{E}_{s_1\sim\mu_1,a_1\sim\widetilde{\pi}}[\sum\!\!\!\!\!\!\!\int_{s_2}\Big(\mathbb{P}_{T,1}(s_2|s_1,a_1,\mu^\pi_{M,1})-\mathbb{P}_{\widetilde{T},1}(s_2|s_1,a_1,\mu^\pi_{\widetilde{M},1})\Big)V^{\widetilde{\pi}}_{\widetilde{M},2}(s_2;\boldsymbol{\mu}^\pi_{\widetilde{M}})]|\\
&+|\mathbb{E}_{s_1\sim\mu_1,a_1\sim\widetilde{\pi}}[\sum\!\!\!\!\!\!\!\int_{s_2}\mathbb{P}_{T,1}(s_2|s_1,a_1,\mu^\pi_{M,1})\Big(V^{\widetilde{\pi}}_{M,2}(s_2;\boldsymbol{\mu}^\pi_M)-V^{\widetilde{\pi}}_{\widetilde{M},2}(s_2;\boldsymbol{\mu}^\pi_{\widetilde{M}})\Big)]|\\
\leq&L_r\|\mu^\pi_{M,1}-\mu^\pi_{\widetilde{M},1}\|_{\mathbb{TV}}+\mathbb{E}_{s_1\sim\mu_1,a_1\sim\widetilde{\pi}}[\|\mathbb{P}_{T,1}(\cdot|s_1,a_1,\mu^\pi_{M,1})-\mathbb{P}_{\widetilde{T},1}(\cdot|s_1,a_1,\mu^\pi_{\widetilde{M},1})\|_{\mathbb{TV}}]\\
&+|\mathbb{E}_{s_1\sim\mu_1,a_1\sim\widetilde{\pi},s_2\sim\mathbb{P}_{T,1}(\cdot|s_1,a_1,\mu^\pi_M)}[V^{\widetilde{\pi}}_{M,2}(s_2;\boldsymbol{\mu}^\pi_M)-V^{\widetilde{\pi}}_{\widetilde{M},2}(s_2;\boldsymbol{\mu}^\pi_{\widetilde{M}})]|\\
\leq&\sum_{h=1}^{H}L_r\|\mu^\pi_{M,h}-\mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}}+\mathbb{E}_{\widetilde{\pi},M|\boldsymbol{\mu}^\pi_M}[\sum_{h=1}^{H}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h})-\mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}].
\end{aligned}
\tag{19}
$$

we finish the proof by applying Corollary D.2. □

Theorem D.5 (Result for MFC; Full Version of Thm. 4.1)**.** *Under Assump.A, B, by running Alg. 1 with the MFC branch, after consuming HK trajectories in Alg. 1 and additional $O(\frac{1}{\varepsilon^2}\log^2\frac{1}{\delta})$ trajectories in the policy selection process in Alg. 3, where K is set to*

$$K = \widetilde{O}\Big((1+L_rH)^2(1+L_TH)^2\Big(\frac{(1+L_T)^H-1}{L_T}\Big)^2\frac{dimE_\alpha(\mathcal{M},\varepsilon_0)}{\varepsilon^2}\Big)$$

with

$$\varepsilon_0 = O(\frac{L_T\varepsilon}{\alpha H(1+L_rH)(1+L_TH)((1+L_T)^H-1)}).$$

or set to the following under additional Assump. D:

$$K = \widetilde{O}\Big((1+L_rH)^2(1+L_TH)^2\Big(1+\frac{L_T}{1-L_\Gamma}\Big)^2\frac{dimE_\alpha(\mathcal{M},\varepsilon_0)}{\varepsilon^2}\Big),$$

with

$$\varepsilon_0 = O(\frac{\varepsilon}{\alpha H(1+L_rH)(1+L_TH)}(1+\frac{L_T}{1-L_\Gamma})^{-1}).$$

with probability at least $1-5\delta$, we have $\mathcal{E}_{Opt}(\widehat{\pi}^*_{Opt}) \leq \varepsilon$.

Proof. On the event of Thm. 4.2, by Lem. 4.5, we have:

$$\begin{aligned}\sum_{k=1}^K \mathcal{E}_{\text{Opt}}(\pi^{k+1}) \leq& \sum_{k=1}^K J_{M^{k+1}}(\pi^{k+1}) - J_{M^*}(\pi^{k+1}) \qquad (M^*\in\widehat{\mathcal{M}}^{k+1})\\ \leq& \sum_{k=1}^K \mathbb{E}_{\pi^{k+1},M^*}[\sum_{h=1}^H (1+L_rH)\|\mathbb{P}_{T^*,h}(\cdot|s_h,a_h,\mu^{\pi^{k+1}}_{M^*,h}) - \mathbb{P}_{T^{k+1},h}(\cdot|s_h,a_h,\mu^{\pi^{k+1}}_{M^{k+1},h})\|_{\mathbb{TV}}].\end{aligned}$$

Next, by applying Thm. 4.3 and Thm. C.4, w.p. $1-3\delta$, for any $\varepsilon_0>0$, we have:

$$\sum_{k=1}^K \mathcal{E}_{\text{Opt}}(\pi^{k+1}) = O\Big((1+L_TH)(1+L_rH)\frac{(1+L_T)^H-1}{L_T}\Big(\sqrt{K\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}}+\alpha HK\varepsilon_0\Big)\Big).$$

Now take a look at Alg. 3, for each $n\in[N]$, by Markov inequality, with probability at least $\frac{2}{3}$:

$$\mathcal{E}_{\text{Opt}}(\pi^{k_n}) = J_{M^*}(\pi^*_{\text{Opt}}) - J_{M^*}(\pi^{k_n}) \tag{20}$$

$$\leq 3\cdot\frac{1}{K}\cdot O\Big((1+L_TH)(1+L_rH)\frac{(1+L_T)^H-1}{L_T}\Big(\sqrt{K\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}}+\alpha HK\varepsilon_0\Big)\Big). \tag{21}$$

$$= O\Big((1+L_TH)(1+L_rH)\frac{(1+L_T)^H-1}{L_T}\Big(\sqrt{\frac{1}{K}\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}}+\alpha H\varepsilon_0\Big)\Big). \tag{22}$$

Since $\pi^{k_1},...,\pi^{k_N}$ are i.i.d. randomly selected, by choosing:

$$K = \widetilde{O}\Big((1+L_TH)^2(1+L_rH)^2\Big(\frac{(1+L_T)^H-1}{L_T}\Big)^2\frac{\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)}{\varepsilon^2}\Big)$$

with $\varepsilon_0 = O(\frac{L_T\varepsilon}{\alpha H(1+L_TH)(1+L_rH)((1+L_T)^H-1)})$, to make sure the RHS of Eq. (22) is less than $\frac{\varepsilon}{2}$. Therefore, in Alg. 3, with probability $1-\delta$, we have

$$\exists n\in[N],\quad \mathcal{E}_{\text{Opt}}(\pi^{k_n}) \leq \frac{\varepsilon}{2}.$$

Then, by Hoeffding inequality, and note that the total return is upper bounded by 1, on good events of concentration, with probability $1-\delta$, we have:

$$\forall n \in [N], \quad |\widehat{J}_{M^*}(\pi^{k_n}) - J_{M^*}(\pi^{k_n})| \leq \frac{\varepsilon}{4}.$$

which implies

$$J_{M^*}(\widehat{\pi}^*_{\text{Opt}}) \geq \max_{n\in[N]} J_{M^*}(\pi^{k_n}) - \frac{\varepsilon}{2} \geq J_{M^*}(\pi^*_{\text{Opt}}) - \varepsilon.$$

Combining all the failure rate together, the above holds with probability at least $1-5\delta$.

The analysis is similar with additional Assump. D, where we have:

$$\sum_{k=1}^K \mathcal{E}_{\text{Opt}}(\pi^{k+1}) = O\Big((1+L_T H)(1+L_r H)(1+\frac{L_T}{1-L_\Gamma})\Big(\sqrt{K\text{dimE}(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}} + \alpha HK\varepsilon_0\Big)\Big),$$

and we should choose

$$K = \widetilde{O}\Big((1+L_T H)^2(1+L_r H)^2\Big(1+\frac{L_T}{1-L_\Gamma}\Big)^2\frac{\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)}{\varepsilon^2}\Big),$$

with $\varepsilon_0 = O(\frac{\varepsilon}{\alpha H(1+L_T H)(1+L_r H)}(1+\frac{L_T}{1-L_\Gamma})^{-1})$. □

D.4 Proofs for RL for Mean-Field Game

Lemma 4.6. *Given two arbitrary model M and $\widetilde{M}$, and two policies π and $\widetilde{\pi}$, we have:*

$$\begin{aligned}|\Delta_M(\widetilde{\pi},\pi) - \Delta_{\widetilde{M}}(\widetilde{\pi},\pi)| \leq & \mathbb{E}_{\widetilde{\pi},M|\boldsymbol{\mu}^\pi_M}[\sum_{h=1}^H \|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}] \\ & + (2L_r H + 1)\mathbb{E}_{\pi,M}[\sum_{h=1}^H \|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]. \qquad (9)\end{aligned}$$

Proof. First of all,

$$\begin{aligned}|\Delta_M(\widetilde{\pi},\pi) - \Delta_{\widetilde{M}}(\widetilde{\pi},\pi)| = & |J_M(\widetilde{\pi};\boldsymbol{\mu}^\pi_M) - J_M(\pi;\boldsymbol{\mu}^\pi_M) - J_{\widetilde{M}}(\widetilde{\pi};\boldsymbol{\mu}^\pi_{\widetilde{M}}) + J_{\widetilde{M}}(\pi;\boldsymbol{\mu}^\pi_{\widetilde{M}})| \\ \leq & |J_M(\widetilde{\pi};\boldsymbol{\mu}^\pi_M) - J_{\widetilde{M}}(\widetilde{\pi};\boldsymbol{\mu}^\pi_{\widetilde{M}})| + |J_M(\pi;\boldsymbol{\mu}^\pi_M) - J_{\widetilde{M}}(\pi;\boldsymbol{\mu}^\pi_{\widetilde{M}})|.\end{aligned}$$

From Eq. (19) of Lem. 4.5, we have:

$$\begin{aligned}& |J_M(\widetilde{\pi};\boldsymbol{\mu}^\pi_M) - J_{\widetilde{M}}(\widetilde{\pi};\boldsymbol{\mu}^\pi_{\widetilde{M}})| \\ \leq & \sum_{h=1}^H L_r\|\mu^\pi_{M,h} - \mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}} + \mathbb{E}_{\widetilde{\pi},M|\boldsymbol{\mu}^\pi_M}[\sum_{h=1}^H \|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}].\end{aligned}$$

By choosing $\widetilde{\pi} = \pi$, the above implies

$$\begin{aligned}& |J_M(\pi;\boldsymbol{\mu}^\pi_M) - J_{\widetilde{M}}(\pi;\boldsymbol{\mu}^\pi_{\widetilde{M}})| \\ \leq & \sum_{h=1}^H L_r\|\mu^\pi_{M,h} - \mu^\pi_{\widetilde{M},h}\|_{\mathbb{TV}} + \mathbb{E}_{\pi,M}[\sum_{h=1}^H \|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h,a_h,\mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}\Big].\end{aligned}$$

Therefore,

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

where we have:

$$\sum_{h=1}^{H} \|\mu^{\pi}_{M,h} - \mu^{\pi}_{\widetilde{M},h}\|_{\mathbb{TV}} \leq H\mathbb{E}_{\pi,M}[\sum_{h=1}^{H} \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu^{\pi}_{M,h}) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h})\|_{\mathbb{TV}}].$$

As a result of Corollary. D.2, and we finish the proof. □

Theorem D.6 (Result for MFG; Full Version of Thm. 4.1)**.** *Under Assump. A and B, by running Alg. 1 with the MFG branch, after consuming $2HK$ trajectories, where K is set to*

$$K = \widetilde{O}\Big((1 + L_T H)^2 (1 + L_r H)^2 \Big(\frac{(1 + L_T)^H - 1}{L_T}\Big)^2 \frac{\mathit{dimE}_{\alpha}(\mathcal{M}, \varepsilon_0)}{\varepsilon^2}\Big),$$

where $\varepsilon_0 = O(\frac{L_T \varepsilon}{\alpha H(1+L_T H)(1+L_r H)((1+L_T)^H - 1)})$; or set to the following with additional Assump. D:

$$K = \widetilde{O}\Big((1 + L_T H)^2 (1 + L_r H)^2 \Big(1 + \frac{L_T}{1 - L_\Gamma}\Big)^2 \frac{\mathit{dimE}_{\alpha}(\mathcal{M}, \varepsilon_0)}{\varepsilon^2}\Big),$$

where $\varepsilon_0 = O(\frac{\varepsilon}{\alpha H(1+L_T H)(1+L_r H)}(1 + \frac{L_T}{1-L_\Gamma})^{-1})$, with probability at least $1 - 5\delta$, we have $\mathcal{E}_{NE}(\widehat{\pi}^*_{NE}) \leq \varepsilon$.

Proof. In the following, we use $\mathcal{E}^M_{\text{NE}}(\pi) := \max_{\widetilde{\pi}} \Delta_M(\widetilde{\pi}, \pi)$ to denote the exploitability in model M. Recall M^{k+1} denotes the model such that π^{k+1} is one of its equilibrium policies satisfying $\mathcal{E}^{M^{k+1}}_{\text{NE}}(\pi^{k+1}) = 0$. On the event in Thm. 4.2, $\forall k \in [K]$, we have $M^* \in \widehat{\mathcal{M}}^k$, which implies

$$\begin{aligned}
\mathcal{E}_{\text{NE}}(\pi^{k+1}) \leq & \mathcal{E}^{\widetilde{M}^{k+1}}_{\text{NE}}(\pi^{k+1}) \\
= & \Delta_{\widetilde{M}^{k+1}}(\widetilde{\pi}^{k+1}, \pi^{k+1}) \\
\leq & \Delta_{\widetilde{M}^{k+1}}(\widetilde{\pi}^{k+1}, \pi^{k+1}) - \Delta_{M^{k+1}}(\widetilde{\pi}^{k+1}, \pi^{k+1}) \\
& \qquad\qquad (\pi^{k+1} \text{ is an equilibrium policy of } M^{k+1} \text{ so } \Delta_{M^{k+1}}(\widetilde{\pi}^{k+1}, \pi^{k+1}) \leq 0) \\
\leq & |\Delta_{\widetilde{M}^{k+1}}(\widetilde{\pi}^{k+1}, \pi^{k+1}) - \Delta_{M^*}(\widetilde{\pi}^{k+1}, \pi^{k+1})| + |\Delta_{M^*}(\widetilde{\pi}^{k+1}, \pi^{k+1}) - \Delta_{M^{k+1}}(\widetilde{\pi}^{k+1}, \pi^{k+1})|.
\end{aligned}$$

By applying Lem. 4.6, Coro. D.2, and Thm. C.4, under Assump. B, we have:

$$\begin{aligned}
\sum_{k=1}^{K} \mathcal{E}_{\text{NE}}(\pi^{k+1}) \leq & \sum_{k=1}^{K} \mathcal{E}^{\widetilde{M}^{k+1}}_{\text{NE}}(\pi^{k+1}) \\
= & O\Big((1 + L_T H)(1 + L_r H)\frac{(1 + L_T)^H - 1}{L_T}\Big(\sqrt{K\text{dimE}_{\alpha}(\mathcal{M}, \varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}} + \alpha KH\varepsilon_0\Big)\Big).
\end{aligned}$$

For the choice of $\widehat{\pi}^*_{\text{NE}}$, since

$$\mathcal{E}_{\text{NE}}(\widehat{\pi}^*_{\text{NE}}) \leq \min_{k \in [K]} \mathcal{E}^{\widetilde{M}^{k+1}}_{\text{NE}}(\pi^{k+1}) \leq \frac{1}{K}\sum_{k=1}^{K} \mathcal{E}^{\widetilde{M}^{k+1}}_{\text{NE}}(\pi^{k+1}),$$

$\mathcal{E}_{\text{NE}}(\widehat{\pi}^*_{\text{NE}}) \leq \varepsilon$ can be ensured by:

$$K = \widetilde{O}\Big((1 + L_T H)^2 (1 + L_r H)^2 \Big(\frac{(1 + L_T)^H - 1}{L_T}\Big)^2 \frac{\text{dimE}_{\alpha}(\mathcal{M}, \varepsilon_0)}{\varepsilon^2}\Big),$$

where $\varepsilon_0 = O(\frac{L_T \varepsilon}{\alpha H(1+L_T H)(1+L_r H)((1+L_T)^H - 1)})$.

Given additional Assump. D, we have:

$$\begin{aligned}
\sum_{k=1}^{K} \mathcal{E}_{\text{NE}}(\pi^{k+1}) \leq & \sum_{k=1}^{K} \mathcal{E}^{\widetilde{M}^{k+1}}_{\text{NE}}(\pi^{k+1}) \\
= & O\Big((1 + L_T H)(1 + L_r H)(1 + \frac{1}{1 - L_\Gamma})\Big(\sqrt{K\text{dimE}_{\alpha}(\mathcal{M}, \varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}} + \alpha KH\varepsilon_0\Big)\Big).
\end{aligned}$$

$\mathcal{E}_{\text{NE}}(\widehat{\pi}^*_{\text{NE}}) \leq \varepsilon$ can be ensured by

$$K = \widetilde{O}\Big((1 + L_T H)^2 (1 + L_r H)^2 \Big(1 + \frac{L_T}{1 - L_\Gamma}\Big)^2 \frac{\text{dimE}_\alpha(\mathcal{M}, \varepsilon_0)}{\varepsilon^2}\Big),$$

where $\varepsilon_0 = O(\frac{\varepsilon}{\alpha H(1+L_T H)(1+L_r H)}(1 + \frac{L_T}{1-L_\Gamma})^{-1})$.

We finish the proof by noting that the total failure rate would be $1 - 3\delta$, and the total sample complexity would be $2HK$. □

E Questions Concerning Existence and Imposed Conditions

In this section, we analyze the existence of MFG-NE in the game described and discuss when the presented conditions might be satisfied. For clarity in notation, we fix the model $M = (\{\mathbb{P}_{T,h}\}_{h=1}^H, \{\mathbb{P}_{r,h}\}_{h=1}^H)$ and the initial distribution μ_1, and also for simplicity denote the deterministic expected rewards

$$r_h(s, a, \mu) := \mathbb{E}_{r \sim \mathbb{P}_{r,h}(\cdot|s,a,\mu)}\left[r\right],$$

since the probabilistic distribution of rewards will not be significant for existence results. In the presented MFG-NE problem, the goal is to find a sequence of policies $\pi := \{\pi_h\}_{h=1}^H$ and a sequence of population distributions $\boldsymbol{\mu} = \{\mu_h\}_{h=1}^H$ such that

$$\textbf{Consistency: } \mu_{h+1} = \Gamma_{pop,h}(\mu_h, \pi_h), \forall h = 1, \dots, H-1,$$
$$\textbf{Optimality: } J_M(\pi, \boldsymbol{\mu}) = \max_{\pi'} J_M(\pi', \boldsymbol{\mu})$$

where μ_1 is fixed and for any $\boldsymbol{\mu} = \{\mu_h\}_{h=1}^H$, $\pi := \{\pi_h\}_{h=1}^H$, with $\mu_h \in \Delta(\mathcal{S}_h)$ and $\pi_h \in \Pi_h := \{\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A}_h)\}$. We define:

$$\Gamma_{pop,h}(\mu_h, \pi_h) := \mathop{\sum\!\!\!\!\!\!\!\int}_{s_h \in \mathcal{S}_h} \mathop{\sum\!\!\!\!\!\!\!\int}_{a_h \in \mathcal{A}_h} \mu_h(s_h)\pi_h(a_h|s_h)\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h),$$

$$J_M(\pi, \boldsymbol{\mu}) := \mathbb{E}\left[\sum_{h=1}^H r_h(s_h, a_h, \mu_h)\middle|_{\substack{s_1 \sim \mu_1,\ a_h \sim \pi_h \\ s_{h+1} \sim \mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h),}} \ \forall h \geq 1\right].$$

As a general strategy, we formulate in this section the two MFG-NE conditions above as fixed point problems. Throughout this section, we will assume the following:

Assumption E (Continuous rewards and dynamics)**.** For each $h \in [H]$, $(s_h, a_h, s_{h+1}) \in \mathcal{S}_h \times \mathcal{A}_h \times \mathcal{S}_{h+1}$, the mappings

$$\mu \to r_h(s_h, a_h, \mu); \quad \mu \to \mathbb{P}_{T,h}(s_{h+1}|s_h, a_h, \mu)$$

are continuous, where $\Delta(\mathcal{S})$ is equipped with the total variation distance $\mathbb{TV}$.

E.1 MFG-NE as a Fixed Point

We use the standard definition of Q-value functions on finite horizon MF-MDPs, for any $\bar{h}, s, a, \pi, \boldsymbol{\mu}$ given by

$$Q_{\bar{h}}^{\pi}(s_{\bar{h}}, a_{\bar{h}}, \boldsymbol{\mu}) := \mathbb{E}\left[\sum_{h=\bar{h}}^H r_h(s_h, a_h, \mu_h)\middle| a_h \sim \pi_h(s_h), s_{h+1} \sim \mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h),\ \forall\ h > \bar{h}\right]. \tag{23}$$

Observe that the set of policies and $\Delta(\mathcal{S})$ are both convex and closed sets (in fact, polytopes), given by $\{\Delta(\mathcal{S}_h)\}_{h\in[H]}, \{\Pi_h\}_{h\in[H]}$. We equip these sets with the metrics

$$\forall \pi, \pi' \in \{\Pi_h\}_{h\in[H]}, \quad d_1(\pi, \pi') := \sup_h \|\pi_h - \pi'_h\|_2$$
$$\forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \{\Delta(\mathcal{S}_h)\}_{h\in[H]}, \quad d_2(\boldsymbol{\mu}, \boldsymbol{\mu}') := \sup_h \|\mu_h - \mu'_h\|_2.$$

We also define the operators $\Gamma_{pop} : \{\Pi_h\}_{h\in[H]} \rightarrow \{\Delta(\mathcal{S}_h)\}_{h\in[H]}$ and $\Gamma_{pp} : \{\Pi_h\}_{h\in[H]} \times \{\Delta(\mathcal{S}_h)\}_{h\in[H]} \rightarrow \{\Pi_h\}_{h\in[H]}$ as

$$\Gamma_{pop}(\pi) := \{\mu_1\} \cup \{\mu_{h+1} := \underbrace{(\Gamma_{pop}(\pi_h, \dots \Gamma_{pop,2}(\pi_2, \Gamma_{pop,1}(\pi_1, \mu_1)))}_{\text{from 1 to } h}\}_{h=1}^{H-1},$$

$$\Gamma_{pp}(\pi, \boldsymbol{\mu}) := \{\pi'_h(\cdot|s_h) := \arg\max_{u\in\Delta_{\mathcal{A}}} Q_h^\pi(s_h, \cdot, \boldsymbol{\mu})^\top u - \|\pi_h(\cdot|s_h) - u\|_2^2\}_{h=1}^H,$$

where Q_h^π is the Q-value function defined in Eq. (23). The motivation for these operators is given by the following lemma:

Lemma E.1 (MFG-NE as fixed point)**.** *The tuple $\pi^*, \boldsymbol{\mu}^*$ is a MFG-NE if and only if the following conditions hold:*

1. $\pi^* = \Gamma_{pp}(\pi^*, \Gamma_{pop}(\pi^*))$*, that is, π^* is a fixed point of* $\Gamma_{NE}(\cdot) := \Gamma_{pp}(\cdot, \Gamma_{pop}(\cdot))$.
2. $\boldsymbol{\mu}^* = \Gamma_{pop}(\pi^*)$.

Proof. First, assume $(\pi^*, \boldsymbol{\mu}^*)$ is a MFG-NE, i.e., it satisfies the consistency and optimality conditions. By consistency, we have $\Gamma_{pop}(\pi^*) = \boldsymbol{\mu}^*$, and since this implies $\Gamma_{pp}(\pi^*, \boldsymbol{\mu}^*) = \pi^*$, the optimality condition implies for each h, s,

$$\pi_h^*(\cdot|s) = \arg\max_{u\in\Delta_{\mathcal{A}}} Q_h^{\pi^*}(s, \cdot, \boldsymbol{\mu}^*)^\top u.$$

which implies that

$$\pi_h^*(\cdot|s) = \arg\max_{u\in\Delta_{\mathcal{A}}} Q_h^{\pi^*}(s, \cdot, \boldsymbol{\mu}^*)^\top u - \|\pi_h^*(\cdot|s) - u\|_2^2,$$

that is, $\Gamma_{NE}(\pi^*) = \pi^*$.

Conversely, assume $\pi^* = \Gamma_{NE}(\pi^*)$, that is, π^* is a fixed point of the operator Γ_{NE}. We claim that $(\pi^*, \boldsymbol{\mu}^* = \Gamma_{pop}(\pi^*))$ is a MFG-NE. For this pair, the consistency condition is satisfied by definition, and the fixed point condition reduces to $\Gamma_{pp}(\pi^*, \boldsymbol{\mu}^*) = \pi^*$. Writing out the definition of the Γ_{pp} operator, we obtain for each h and s_h,

$$\pi_h^*(\cdot|s) = \arg\max_{u\in\Delta_{\mathcal{A}}} Q_h^{\pi^*}(s, \cdot, \boldsymbol{\mu}^*)^\top u - \|\pi_h^*(\cdot|s) - u\|_2^2,$$
$$\pi_h^*(\cdot|s) = \arg\max_{u\in\Delta_{\mathcal{A}}} Q_h^{\pi^*}(s, \cdot, \boldsymbol{\mu}^*)^\top u,$$

by the first-order optimality conditions of the term $Q_h^{\pi^*}(s, \cdot, \boldsymbol{\mu}^*)^\top u - \|\pi_h(\cdot|s) - u\|_2^2$. Then, by the optimality conditions of MDPs, π^* is also the optimal policy with respect to $\boldsymbol{\mu}^*$, that is, $J_M(\pi^*, \boldsymbol{\mu}^*) = \max_{\pi'} J_M(\pi', \boldsymbol{\mu}^*)$. □

In the lemma above, the second condition is trivial to satisfy/compute once π^* is known, hence the primary challenge will be in proving that the map Γ_{NE} admits a fixed point.

E.2 Existence of MFG-NE

We use the Brower fixed point method to prove the existence of a MFG-NE, and Assump. E is sufficient. The strategy will be to show that Γ_{NE} is a continuous function on the compact and convex policy/population distribution space.

We will prove several continuity results, in order to be able to apply Brouwer's fixed point theorem.

Lemma E.2 (Continuity of Q_h^π)**.** *For any s, a, h, the map*

$$\pi, \boldsymbol{\mu} \rightarrow Q_h^\pi(s, a, \boldsymbol{\mu}) \in \mathbb{R}$$

is continuous.

Proof. The proof follows from the fact that Q_h^π is a function of sum and multiplications of continuous functions of the policies and population distributions $\{\pi_h\}_{h\in[H]}, \{\mu_h\}_{h\in[H]}$. The compositions, additions and multiplications of continuous functions are continuous. □

For the next continuity result, we will need the following well-known Fenchel conjugate definition and duality.

Definition E.3 (Fenchel conjugate)**.** Assume that $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a convex function, with domain $\mathcal{X} \subset \mathbb{R}^d$. The Fenchel conjugate $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is defined as

$$f^*(y) = \sup_{x \in \mathcal{X}} \langle x, y \rangle - f(x).$$

For further details regarding the Fenchel conjugate, see [44]. The Fenchel conjugate is useful due to the following well-known duality result.

Lemma E.4. *Assume that $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is differentiable and τ-weakly convex and has domain $\mathcal{X} \subset \mathbb{R}^d$. Then,*

1. *f^* is differentiable on $\mathbb{R}^d$,*
2. *$\nabla f^*(y) = \arg\max_{x \in \mathcal{X}} \langle x, y \rangle - f(x)$,*
3. *f^* is $\frac{1}{\tau}$-smooth with respect to $\| \cdot \|_2$, i.e., $\|\nabla f^*(y) - \nabla f^*(y')\| \leq \frac{1}{\tau} \|y - y'\|_2, \forall y, y' \in \mathbb{R}^d$.*

Proof. See Lemma 15 of [54] or Lemma 6.1.2 of [44]. □

Finally, we will also need the non-expansiveness of the proximal point operator, presented below.

Lemma E.5 (Proximal operator is non-expansive [47])**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact convex set, and $f : \mathcal{X} \to \mathbb{R}$ be a convex function. The proximal map $\operatorname{prox}_f : \mathcal{X} \to \mathcal{X}$ defined by*

$$\operatorname{prox}_f(x) := \arg\min_{y \in \mathcal{X}} f(y) + \|x - y\|_2^2$$

is non-expansive (hence continuous).

With the presented tools, we can prove the following statement.

Lemma E.6 (Continuity of $\Gamma_{pop}, \Gamma_{pp}$)**.** *With the metrics d_1, d_2, the operators $\Gamma_{pop}, \Gamma_{pp}$ are Lipschitz continuous mappings.*

Proof. The continuity of Γ_{pop} w.r.t. π is straightforward by definition, as multiplications and additions of continuous functions are continuous.

For the continuity of Γ_{pp}, we can either explicitly write the solution of the $\arg\max$ problem in terms of an affine function and a projection of terms $Q_h^{\pi_h}, \pi_h$, or more generally use Fenchel duality combined with the non-expansiveness of the proximal point operator. By Lemma E.5, the map

$$u \to \arg\max_{u' \in \Delta_{\mathcal{A}}} q^\top u' - \|u - u'\|_2^2 = -\operatorname{prox}_{-q^\top(\cdot)}(u)$$

is a continuous map for any $q \in \mathbb{R}^{|\mathcal{A}|}$. Similarly, by Lemma E.4, the map

$$q \to \arg\max_{u \in \Delta_{\mathcal{A}}} q^\top u - \|u - u_0\|_2^2$$

is differentiable hence continuous for any $u_0 \in \Delta_{\mathcal{A}}$, as the map $\|u_0 - \cdot\|_2^2$ is weakly convex. By the continuity of Q_h^π (see Lemma E.2), we can conclude that Γ_{pp} is also a continuous map, as it is the composition of continuous functions. □

With this continuity characterization, we invoke Brouwer's fixed point theorem to prove existence.

Proposition E.7 (Existence of MFG-NE; Formal Version of Prop. 2.1)**.** *Under Assump. E (which is implied by Assump. B), the map Γ_{NE} has a fixed point in the set $\{\Pi_h\}_{h \in [H]}$, that is, there exists a π^* such that $\Gamma_{NE}(\pi^*) = \pi^*$, and the tuple $(\pi^*, \Gamma_{pop}(\pi^*))$ is a MFG-NE.*

Proof. With the continuity of $\Gamma_{pop}, \Gamma_{pp}$, the know that the composition Γ_{NE} is continuous. It maps the closed, convex polytope $\{\Pi_h\}_{h \in [H]}$ to a subset of itself, hence by Brouwers fixed point theorem it must admit a fixed point. By Lemma E.1, this fixed point must constitute a MFG-NE. □

F Proofs for Lower Bound for Mean-Field Control in Tabular Setting

We first introduce the notion of Strong Query Model (SQM), which is a strictly stronger notion than both GM and DCP, since it can return the entire conditional distribution for any policy or arbitrary density sequence. Our lower bound will be established based on SQM, which directly implies an lower bound for GM or DCP setting.

Definition F.1 (Strong Query Model (SQM))**.** The Strong Generative Model (abbr. SGM) either can be queried by a policy π and return the entire transition distribution of the true model conditioning on $\boldsymbol{\mu}^\pi_{M^*}$, i.e.: $\{\mathbb{P}_{T^*,h}(\cdot|\cdot,\cdot,\mu^\pi_{M^*,h})\}_{h\in[H]}$; or can be queried by an arbitrary sequence of density $\boldsymbol{\mu} := \{\mu_1, ..., \mu_H\}$ and return the entire transition distribution of the true model conditioning on $\boldsymbol{\mu}$, i.e., $\{\mathbb{P}_{T^*,h}(\cdot|\cdot,\cdot,\mu_h)\}_{h\in[H]}$.

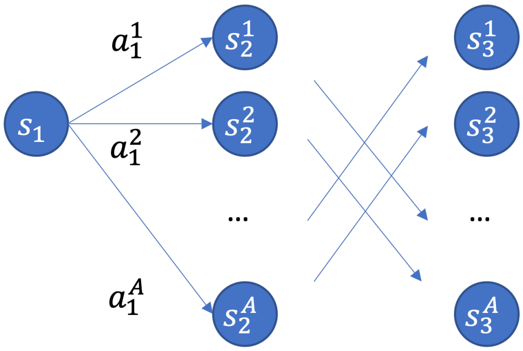

Figure 2: Construction of Lower Bound

Theorem 5.2. *[Exponential Lower Bound for MFC] Given arbitrary $L_T > 0$ and $d \geq 2$, 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 $\varepsilon \leq \frac{L_T}{d+1}$, 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 $K \leq |\mathcal{M}|/2 - 1$, the probability that Alg produces an ε-near-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. 2. 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:

$$\begin{aligned}
&\Pr(M = \bar{M}) + \sum_{n\in[N]} \Pr(\{M = M^n\} \cap \{\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})\}) \\
&\hspace{18em} \text{(Each policy is optimal in } \bar{M}.) \\
=& \frac{1}{|\mathcal{M}|} + \frac{1}{|\mathcal{M}|} \sum_{n\in[N]} \Pr_{M^n,\text{Alg}}(\text{Alg}(K) \in \Pi(\mu^n, \frac{4\varepsilon}{L_T})) \\
\leq& \frac{K+1}{|\mathcal{M}|}.
\end{aligned}$$

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

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

□

Algorithm 4: Bridge Model Construction

```
1  Input: Model Class $\mathcal{M}$; Integer $N$.
2  for $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$ do
3      for $\mu_h \in \mathcal{U}_N$ do
4          $M^{\text{Central}}_{s_h,a_h,\mu_h} \leftarrow \arg\max_{M\in\mathcal{M}} \mathcal{B}_{\mathcal{M}}(M;\bar{\varepsilon})[s_h,a_h,\mu_h]$.
5          $f_{s_h,a_h}(\mu_h) \leftarrow \mathbb{P}_{T^{\text{Central}}_{s_h,a_h,\mu_h},h}(\cdot|s_h,a_h,\mu_h)$.
6          // Here $\mathbb{P}_{T^{\text{Central}}_{s_h,a_h,\mu_h}}$ denotes the transition function of $M^{\text{Central}}_{s_h,a_h,\mu_h}$.
7      end
8      Construct transition for $s_h, a_h$ by:
           $\forall \mu'_h \in \Delta(\mathcal{S}_h), \quad \mathbb{P}_{T_{\text{Br}},h}(\cdot|s_h,a_h,\mu'_h) := \frac{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+ f_{s_h,a_h}(\mu_h)}{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+}.$
9  end
10 Define $M_{\text{Br}}$ to be the model with transitoin $\{\mathbb{P}_{T_{\text{Br}},h}\}_{h\in[H]}$.
11 Compute an equilibrium policy $\pi^{\text{Br}}_{\text{NE}}$ for $M_{\text{Br}}$ and its induced density in $M_{\text{Br}}$ denoted as $\boldsymbol{\mu}^{\text{NE}}_{\text{Br}}$.
12 return $M_{\text{Br}}$, $\pi^{\text{Br}}_{\text{NE}}$, $\boldsymbol{\mu}^{\text{NE}}_{\text{Br}}$.
```

G Learning Mean Field Game with Generative Model in Tabular Setting

We first introduce the algorithm for bridge model construction, where $[x]^+ = \max\{0, x\}$. In this section, given an integer N, we define the set:

$$\mathcal{U}_N := \{\mu = (\frac{x_1}{N}, \frac{x_2}{N}..., \frac{x_S}{N}) \in \mathbb{R}^S | \forall i \in [S],\ x_i \in \mathbb{N}; \sum_{i=1}^S x_1 = N\} \tag{24}$$

First we show that $\mathcal{U}_N$ forms an $\frac{S}{N}$-cover for the density space.

Lemma G.1. *For any $\mu_h \in \Delta(\mathcal{S}_h)$, there exists at least one $\mu \in \mathcal{U}_N$, s.t. $\|\mu_h(s_h) - \mu(s_h)\|_1 \leq \frac{S}{N}$.*

Proof. We make the proof by construction, and we only consider the case when $\mu_h \notin \mathcal{U}_N$. First of all, we define μ_h^+ to be the density vector, such that, $\mu_h(s_h)^+ := \lfloor \mu_h(s_h) \cdot N \rfloor / N$ for any $s_h \in \mathcal{S}_h$. We should have $\mu_h(s_h) - \mu_h^+(s_h) \geq 0$ for any s_h, and $\sum_{s_h} \mu_h(s_h) - \mu_h^+(s_h) = i/N$ for some $i \in \{1, 2, ..., S-1\}$ since $\mu_h \notin \mathcal{U}_N$. Then, we can construct $\bar{\mu}_h$ by assigning $\bar{\mu}_h(s_h) \leftarrow \mu_h^+(s_h) + \frac{1}{N}$ for arbitrary i states $s_h \in \mathcal{S}_h$. Easy to check $\bar{\mu}_h \in \mathcal{U}_N$, and $\|\bar{\mu}_h - \mu_h\|_1 \leq \sum_{s_h} \frac{1}{N} = \frac{S}{N}$. □

Lemma G.2. *When the `Else`-branch in Line 7 is activated, for any $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$, and any $\mu_h^1, \mu_h^2 \in \Delta(\mathcal{S}_h)$, there exists at least one model $M \in \mathcal{M}$ s.t.,*

$$M \in \mathcal{B}_{\mathcal{M}}(M^{Central}_{s_h,a_h,\mu_h^1};\bar{\varepsilon})[s_h,a_h,\mu_h^1] \cap \mathcal{B}_{\mathcal{M}}(M^{Central}_{s_h,a_h,\mu_h^2};\bar{\varepsilon})[s_h,a_h,\mu_h^2],$$

which further implies:

$$\|\mathbb{P}_{T^{Central}_{s_h,a_h,\mu_h^1},h}(\cdot|s_h,a_h,\mu_h^1) - \mathbb{P}_{T^{Central}_{s_h,a_h,\mu_h^2},h}(\cdot|s_h,a_h,\mu_h^2)\|_1 \leq \bar{\varepsilon} + L_T\|\mu_h^1 - \mu_h^2\|_1.$$

Proof. When the `Else`-branch in Line 7 is activated, we have:

$$|\mathcal{B}_{\mathcal{M}}(M^{\text{Central}}_{s_h,a_h,\mu_h^1};\bar{\varepsilon})[s_h,a_h,\mu_h^1]| + |\mathcal{B}_{\mathcal{M}}(M^{\text{Central}}_{s_h,a_h,\mu_h^2};\bar{\varepsilon})[s_h,a_h,\mu_h^2]| > |\mathcal{M}|.$$

which implies that those two sets share at least one common model denoted as M. Let's use $\mathbb{P}_T$ to denote the transition function of M, we have:

$$\begin{aligned}\|\mathbb{P}_{T^{\text{Central}}_{s_h,a_h,\mu_h^1},h}(\cdot|s_h,a_h,\mu_h^1) - \mathbb{P}_{T^{\text{Central}}_{s_h,a_h,\mu_h^2},h}(\cdot|s_h,a_h,\mu_h^2)\|_1 \leq & 2\bar{\varepsilon} + \|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu_h^1) - \mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu_h^2)\|_1 \\ & \text{(Definition of } \mathcal{B}_{\mathcal{M}}) \\ \leq & 2\bar{\varepsilon} + L_T\|\mu_h^1 - \mu_h^2\|_1. \\ & \text{(by Assump. B)}\end{aligned}$$

□

Lemma 5.4. *[Implication of Local Alignment in MFG] Given a model M with transition $\mathbb{P}_T$, suppose M and M^* are locally aligned at policy π w.r.t. the density induced in M, i.e. $\forall h$, $\mathbb{P}_{T^*,h}(\cdot|\cdot,\cdot,\mu^\pi_{M,h}) = \mathbb{P}_{T,h}(\cdot|\cdot,\cdot,\mu^\pi_{M,h})$, if π is a NE in M, then it must be a NE in M^*.*

Proof. Note that $\mu^\pi_{M,1} = \mu^\pi_{M^*,1} = \mu_1$, by inducation, for any $h \in [H]$, we have:

$$\begin{aligned}\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h}) =& \mathbb{P}_{T^*,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M^*,h}) \\ \mu^\pi_{M,h+1}(s_{h+1}) =& \sum_{s_h,a_h} \pi_h(a_h|s_h)\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h}) \\ =& \sum_{s_h,a_h} \pi_h(a_h|s_h)\mathbb{P}_{T^*,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M^*,h}) \\ =& \mu^\pi_{M^*,h+1}(s_{h+1}).\end{aligned}$$

Therefore, if π is the NE of M, for any $\widetilde{\pi}$, we have:

$$0 \leq J_M(\pi;\boldsymbol{\mu}^\pi_M) - J_M(\widetilde{\pi};\boldsymbol{\mu}^\pi_M) = J_{M^*}(\pi;\boldsymbol{\mu}^\pi_{M^*}) - J_{M^*}(\widetilde{\pi};\boldsymbol{\mu}^\pi_{M^*}),$$

which implies π is also the NE of M^*. □

Theorem G.3. *[Formal version of Thm. 5.3] Under Assump. A and B, and hyper-parameter choices $K = \lceil \log_2 |\mathcal{M}| \rceil$, $\bar{\varepsilon} = \frac{\widetilde{\varepsilon}}{10}, N = \frac{\widetilde{\varepsilon}}{10 L_T S}, \bar{N} = O(\frac{S^2}{\bar{\varepsilon}^2} \log \frac{SAHK}{\delta}), \widetilde{N} = O(\frac{S^2}{\widetilde{\varepsilon}^2} \log \frac{SAHK}{\delta})$, with probability $1-\delta$, by choosing*

$$\widetilde{\varepsilon} = \frac{1}{H}(3 + 2(L_r + L_T) \cdot \frac{(1+L_T)^H - 1}{L_T})^{-1}\varepsilon,$$

and consuming number of queries to GM at most:

$$K \cdot (\bar{N} + SAH\widetilde{N}) = O(\frac{S^3 AH \log}{\varepsilon^2}(1 + (L_r + L_T)\frac{(1+L_T)^H - 1}{L_T})^2 \log^2 \frac{SAH|\mathcal{M}|}{\delta});$$

or under additional Assump. D, by choosing

$$\widetilde{\varepsilon} = \frac{1}{H}(3 + 2(L_r + L_T) \cdot \frac{1}{1 - L_\Gamma})^{-1}\varepsilon,$$

and consuming number of queries to GM at most:

$$K \cdot (\bar{N} + SAH\widetilde{N}) = O(\frac{S^3 AH}{\varepsilon^2}(1 + (L_r + L_T)\frac{1}{1 - L_\Gamma})^2 \log^2 \frac{SAH|\mathcal{M}|}{\delta}),$$

Alg. 2 can return us an ε-approximation NE within K iterations.

Proof.

Concentration Events Considering the random variables:

$$X^k := \mathbb{I}[\text{If-branch}] \cdot \Big(\mathbb{P}_{T^*,h}(\cdot|s_h,a_h,\mu_h) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h,a_h,\mu_h)\Big),$$

and

$$Y^k_{h,s_h,a_h} := \mathbb{I}[\text{Else-branch}] \cdot \Big(\mathbb{P}_{T^*,h}(\cdot|s_h,a_h,\mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h,a_h,\mu^{\text{NE},k}_{\text{Br},h})\Big).$$

where $\mathbb{I}[\text{If-branch}]$ equals 1 if the algorithm does not terminate at step k and enters the if-branch, otherwise equals 0; the definition for $\mathbb{I}[\text{Else-branch}]$ is similar. For our choice of $\bar{\varepsilon}$ and $\widetilde{\varepsilon}$, consider the events $\mathcal{E}_{con}$ defined by:

$$\mathcal{E}_{con} := \bigcap_{k=1}^K \Big(\{\|X^k\|_1 \leq \frac{\bar{\varepsilon}}{2}\} \cap \bigcap_{h\in[H],s_h\in\mathcal{S}_h,a_h\in\mathcal{A}_h} \{\|Y^k_{h,s_h,a_h}\|_1 \leq \frac{\widetilde{\varepsilon}}{2}\}\Big).$$

Let's first focus on X^k. Note that the samples from generative models are i.i.d. from the distribution $\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu_h)$, and $\mathbb{P}_{\widehat{T}^*,h}$ is computed by empirical mean. Therefore, no matter $\mathbb{I}$[If-branch] equals 0 or 1, by Hoeffding inequality, w.p. $1-\delta_0$, we have:

$$\|X^k\| = \sum_{s_{h+1}} |\mathbb{I}[\text{If-branch}] \cdot \Big(\mathbb{P}_{T^*,h}(s_{h+1}|s_h, a_h, \mu_h) - \mathbb{P}_{\widehat{T}^*,h}(s_{h+1}|s_h, a_h, \mu_h)\Big)| \leq S\sqrt{\frac{1}{2\bar{N}}\log\frac{S}{\delta_0}}.$$

Similarly, for Y^k, for any fixed h, s_h, a_h, we have $\|Y^k_{h,s_h,a_h}\| \leq S\sqrt{\frac{1}{2\widetilde{N}}\log\frac{S}{\delta_0}}$. By choosing $\delta_0 = \delta/(2SAHK)$ to take the union bound for all steps K and all h, s_h, a_h for Y^k, $\Pr(\mathcal{E}_{con}) \geq 1-\delta$ can be satisfied by:

$$\bar{N} = O(\frac{S^2}{\bar{\varepsilon}^2}\log\frac{SAHK}{\delta}), \quad \widetilde{N} = O(\frac{S^2}{\widetilde{\varepsilon}^2}\log\frac{SAHK}{\delta}).$$

Next, we separately analyze the `If` and `Else` branch in Alg. 2 on the event $\mathcal{E}_{con}$. Obviously, on the event of $\mathcal{E}_{con}$, we have $M^* \in \mathcal{M}^k$ for all $k \in [K]$.

The `If`-branch: Line 3 In this case, there exists $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h, \mu_h \in \Delta(\mathcal{S}_h)$:

$$\max_M |\mathcal{B}_{\mathcal{M}}(M; \bar{\varepsilon})[s_h, a_h, \mu_h]| \leq \frac{1}{2}|\mathcal{M}|.$$

With our choice of $\bar{N}$, we have: $\|\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu_h) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu_h)\|_1 \leq \frac{\bar{\varepsilon}}{2}$. Therefore, for those $M \notin \mathcal{B}_{\mathcal{M}}(M^*; \bar{\varepsilon})[s_h, a_h, \mu_h]$, we must have:

$$\|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu_h)\|_1 > \frac{\bar{\varepsilon}}{2}$$

which implies $|\mathcal{M}^{k+1}| \leq |\mathcal{B}_{\mathcal{M}}(M^*; \bar{\varepsilon})[s_h, a_h, \mu_h]| \leq \frac{1}{2}|\mathcal{M}^k|$.

The `Else-branch`: Line 7 If the `If-branch` is not activated, then for any fixed $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h, \mu_h \in \Delta(\mathcal{S}_h)$, there exists a "Central Model" $M \in \mathcal{M}$, such that,

$$|\mathcal{B}_{\mathcal{M}}(M; \bar{\varepsilon})[s_h, a_h, \mu_h]| > \frac{1}{2}|\mathcal{M}|.$$

Now, consider a fixed $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$. As described in Alg. 4, for each $\mu_h \in \mathcal{U}_N$, we use M_{μ_h} to denote the "Central Model" model has the most number of $\bar{\varepsilon}$-similar models in $\mathcal{M}$ with transtion function $\mathbb{P}_{T^{\text{Central}}_{s_h,a_h,\mu_h}} := \{\mathbb{P}_{T^{\text{Central}}_{s_h,a_h,\mu_h},h}\}_{h\in[H]}$. Besides, we introduce the function f_{s_h,a_h} defined on the density space $\Delta(\mathcal{S}_h)$:

$$\forall \mu_h \in \Delta(\mathcal{S}_h),\ f_{s_h,a_h}(\mu_h) := \mathbb{P}_{T^{\text{Central}}_{s_h,a_h,\mu_h},h}(\cdot|s_h, a_h, \mu_h)$$

In the following, let's use $g_{s_h,a_h}(\cdot)$ as a short note of $\mathbb{P}_{T_{\text{Br}},h}(\cdot|s_h, a_h, \mu'_h)$ constructed from the values of f_{s_h,a_h} on $\mathcal{U}_N$ in Alg. 4:

$$\forall \mu'_h \in \Delta(\mathcal{S}_h), \quad g_{s_h,a_h}(\mu'_h) := \frac{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+ f_{s_h,a_h}(\mu_h)}{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+},$$

By Lem. G.1, there always exists a $\mu_h \in \mathcal{U}_N$ such that $[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+ \geq \frac{S}{N} > 0$, so g is always well-defined. Therefore, we have:

$$\begin{aligned}
\forall \mu'_h \in \Delta(\mathcal{S}_h), \qquad & \|g_{s_h,a_h}(\mu'_h) - f_{s_h,a_h}(\mu'_h)\|_1 \\
&= \|\frac{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+ (f_{s_h,a_h}(\mu_h) - f_{s_h,a_h}(\mu'_h))}{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+}\|_1 \\
&\leq \frac{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+ \|f_{s_h,a_h}(\mu_h) - f_{s_h,a_h}(\mu'_h)\|_1}{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+} \\
&\leq \frac{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+ (2\bar{\varepsilon} + L_T\|\mu_h - \mu'_h\|_1)}{\sum_{\mu_h\in\mathcal{U}_N}[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+} \qquad \text{(Lem. G.2)}
\end{aligned}$$

$$\leq 2\bar{\varepsilon} + 2L_T \frac{S}{N}. \tag{25}$$

where the last step is because $[\frac{2S}{N} - \|\mu'_h - \mu_h\|_1]^+ > 0$ implies $\|\mu'_h - \mu_h\|_1 \leq \frac{2S}{N}$. We can conduct similar discussion for for any $h \in [H], s_h \in \mathcal{S}_h, a_h \in \mathcal{A}_h$.

As stated in Alg. 2, We use M^k_{Br} and π^k_{NE} to denote the bridge model and its equilibrium policy at iteration k if the `Else-branch` is activated. Since both $[\cdot]^+$ and $\|\cdot\|_1$ are continuous, g is continuous w.r.t. $\mu'_h \in \Delta(\mathcal{S}_h)$. By Prop. E.7, the equilibrium policy π^k_{NE} should always exist.

The `Else-If-branch`: Line 10 If Line 10 in Alg. 2 is activated at some h, s_h, a_h, for any $M \in \mathcal{B}_{\mathcal{M}^k}(M_{\mu_h}; \bar{\varepsilon})[s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}]$, where M_{μ_h} denotes the Central Model, we have:

$$\begin{aligned}
&\|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1 \\
\geq&\|\mathbb{P}_{T^k_{\text{Br}},h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1 - \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{T^k_{\text{Br}},h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1 \\
\geq&\widetilde{\varepsilon} - (2\bar{\varepsilon} + 2L_T\frac{S}{N}) \geq \frac{2\widetilde{\varepsilon}}{5} > \frac{\widetilde{\varepsilon}}{2}. && \text{(Eq. (25); Our choice of } N \text{ and } \bar{\varepsilon})
\end{aligned}$$

which implies $|\mathcal{M}^{k+1}| \leq |\mathcal{M}^k| - |\mathcal{B}_{\mathcal{M}^k}(M_{\mu_h}; \bar{\varepsilon})[s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}]| \leq \frac{1}{2}|\mathcal{M}^k|$.

The `Else-Else-branch`: Line 13 If this branch is activated, we must have:

$$\begin{aligned}
&\|\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{T^k_{\text{Br}},h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1 \\
\leq&\|\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1 + \|\mathbb{P}_{\widehat{T}^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{T^k_{\text{Br}},h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1 \\
\leq&2\widetilde{\varepsilon}. && (26)
\end{aligned}$$

In the following, we use $\mathcal{E}^M_{\text{NE}}(\pi) := \max_{\widetilde{\pi}} \Delta_M(\widetilde{\pi}, \pi)$ to denote the exploitability in model M. As a result:

$$\begin{aligned}
&\mathcal{E}_{\text{NE}}(\pi^{\text{Br},k}_{\text{NE}}) \\
=&\mathcal{E}_{\text{NE}}(\pi^{\text{Br},k}_{\text{NE}}) - \mathcal{E}^{M^k_{\text{Br}}}_{\text{NE}}(\pi^{\text{Br},k}_{\text{NE}}) \\
=&\max_\pi J_{M^*}(\pi; \boldsymbol{\mu}^{\text{NE},k}_{M^*}) - J_{M^*}(\pi^{\text{Br},k}_{\text{NE}}; \boldsymbol{\mu}^{\text{NE},k}_{M^*}) - (\max_\pi J_{M^k_{\text{Br}}}(\pi; \boldsymbol{\mu}^{\text{NE},k}_{\text{Br}}) - J_{M^k_{\text{Br}}}(\pi^{\text{Br},k}_{\text{NE}}; \boldsymbol{\mu}^{\text{NE},k}_{\text{Br}})) \\
\leq&2\max_\pi |J_{M^*}(\pi; \boldsymbol{\mu}^{\text{NE},k}_{M^*}) - J_{M^k_{\text{Br}}}(\pi; \boldsymbol{\mu}^{\text{NE},k}_{\text{Br}})| \\
\leq&2\max_\pi \Big(\sum_{h=1}^H L_r\|\mu^{\text{NE},k}_{M^*,h} - \mu^{\text{NE},k}_{\text{Br},h}\|_1 + \mathbb{E}_{\pi, M^k_{\text{Br}}|\boldsymbol{\mu}^{\text{NE},k}_{\text{Br}}}[\sum_{h=1}^H \|\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{M^*,h}) - \mathbb{P}_{T^k_{\text{Br}},h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1]\Big) \\
&&\text{(Eq. (19) in the proof of Lem. 4.5)} \\
\leq&2\max_\pi \Big(\sum_{h=1}^H (L_r + L_T)\|\mu^{\text{NE},k}_{M^*,h} - \mu^{\text{NE},k}_{\text{Br},h}\|_1 + \mathbb{E}_{\pi, M^k_{\text{Br}}|\boldsymbol{\mu}^{\text{NE},k}_{\text{Br}}}[\sum_{h=1}^H \|\mathbb{P}_{T^*,h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h}) - \mathbb{P}_{T^k_{\text{Br}},h}(\cdot|s_h, a_h, \mu^{\text{NE},k}_{\text{Br},h})\|_1]\Big). \\
&&\text{(Assump. B)}
\end{aligned}$$

For any fixed π, the second part can be upper bounded by Eq. (26), while for the first part, by Coro. D.2, under Assump. B, we have:

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

which implies $\mathcal{E}_{\text{NE}}(\pi^{\text{Br},k}_{\text{NE}}) \leq H\widetilde{\varepsilon}(3 + 2(L_r + L_T) \cdot \frac{(1+L_T)^H - 1}{L_T})$. Therefore, by choosing

$$\widetilde{\varepsilon} = \frac{1}{H}(3 + 2(L_r + L_T) \cdot \frac{(1+L_T)^H - 1}{L_T})^{-1}\varepsilon,$$

we can ensure $\pi_{\text{NE}}^{\text{Br},k}$ is an ε-approximation NE for M^*.

Similarly, with additional Assump. D, we have: $\sum_{h=1}^H \|\mu_{M^*,h}^{\text{NE},k} - \mu_{\text{Br},h}^{\text{NE},k}\|_1 \leq 2H\widetilde{\varepsilon}\frac{1}{1-L_\Gamma}$, which implies $\mathcal{E}_{\text{NE}}(\pi_{\text{NE}}^{\text{Br},k}) \leq H\widetilde{\varepsilon}(3 + 2(L_r + L_T) \cdot \frac{1}{1-L_\Gamma})$, and by choosing

$$\widetilde{\varepsilon} = \frac{1}{H}(3 + 2(L_r + L_T) \cdot \frac{1}{1-L_\Gamma})^{-1}\varepsilon,$$

we can also ensure $\pi_{\text{NE}}^{\text{Br},k}$ is an ε-approximation NE for M^*.

Summary From the discussion above, we know that, on the event of $\mathcal{E}_{con}$, either we have $|\mathcal{M}^{k+1}| \leq \frac{1}{2}|\mathcal{M}^k|$, or the algorithm can return an ε-approximate NE. Therefore, by choosing $K = \lceil \log_2 |\mathcal{M}| \rceil$, with probability $1 - \delta$, after consuming number of queries to GM at most:

$$K \cdot (\bar{N} + SAH\widetilde{N}) = O(\frac{S^3AH}{\varepsilon^2}(1 + (L_r + L_T)\frac{(1+L_T)^H - 1}{L_T})^2 \log^2 \frac{SAH|\mathcal{M}|}{\delta}),$$

or under Assump. D, at most

$$K \cdot (\bar{N} + SAH\widetilde{N}) = O(\frac{S^3AH}{\varepsilon^2}(1 + (L_r + L_T)\frac{1}{1-L_\Gamma})^2 \log^2 \frac{SAH|\mathcal{M}|}{\delta}),$$

Alg. 2 can return us an ε-approximation NE. □