# On the Statistical Efficiency of Mean-Field Reinforcement Learning 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 fundamental statistical efficiency of Reinforcement Learning in Mean-Field Control (MFC) and Mean-Field Game (MFG) with general model-based function approximation. We introduce a new concept called Mean-Field Model-Based Eluder Dimension (MF-MBED), which characterizes the inherent complexity of mean-field model classes. We show that a rich family of Mean-Field RL problems exhibits low MF-MBED. Additionally, we propose algorithms based on maximal likelihood estimation, which can return an $\varepsilon$-optimal policy for MFC or an $\varepsilon$-Nash Equilibrium policy for MFG. The overall sample complexity depends only polynomially on MF-MBED, which is potentially much lower than the size of state-action space. Compared with previous works, our results only require the minimal assumptions including realizability and Lipschitz continuity.

## 1 INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) 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 (Shalev-Shwartz et al., 2016), finance (Lee et al., 2007), and robotics control (Ismail et al., 2018). Although MARL has attracted increasing attention in recent RL research (Zhang et al., 2021), when the number of agents is in the hundreds or thousands, solving MARL becomes a challenge. However, in scenarios where agents exhibit symmetry, like humans in crowds or individual cars in the traffic flow, mean-field theory can be employed to approximate the system dynamics, which results in the Mean-Field RL (MFRL) setting. In MFRL, the interaction within large populations is reflected by the dependence on the state density of the transition and reward functions of individual agents. The mean-field model has achieved success in various domains, including economics (Cousin et al., 2011; Angiuli et al., 2021), finance (Cardaliaguet and Lehalle, 2018), industrial engineering (De Paola et al., 2019), etc.

Depending on the objectives, MFRL can be divided into two categories: Mean-Field Control (MFC) and Mean-Field Game (MFG) (Lasry and Lions, 2007; Huang et al., 2006; Bensoussan et al., 2013). 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 tends to deviate. Compared with single-agent RL, MFRL is much more challenging because it requires exploration in the joint space of state, action, and state density, especially given that the density belongs to an infinite and continuous space.

The sample efficiency (referring to the number of samples needed to explore the environment and achieve objectives) in both tabular and more broadly function approximation settings has been extensively examined as one of the fundamental questions in reinforcement learning, particularly in single-agent scenarios. See for example existing work in single-agent (Auer et al., 2008; Azar et al., 2017; Jin et al., 2018; Russo and Van Roy, 2013; Jin et al., 2021; Du et al., 2021; Xie et al., 2022; Foster et al., 2021) and general multi-agent settings (Jin et al., 2023; Bai et al., 2020; Xie et al., 2020; Huang et al., 2021; Wang et al., 2023; Cui et al., 2023). However, the understanding of fundamental sample efficiency in existing MFRL literature is still limited in the following two aspects.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

- **Lack of understanding in fundamental sample efficiency**: Previous works, especially in MFG setting, mainly consider the computational complexity, and therefore, they only focus on settings with strong structural assumptions, such as contractivity (Guo et al., 2019; Xie et al., 2021), monotonicity (Perrin et al., 2020; Pérolat et al., 2022; Elie et al., 2020), or population-independent dynamics (Mahajan, 2021; Geist et al., 2022). As a result, the fundamental sample efficiency under general conditions is still an open problem. On the other hand, those structural assumptions also simplify the exploration process to some extent, so their algorithms can be hardly generalized beyond those assumptions.

- **Lack of understanding in function approximation setting**: Most previous literature only focuses on tabular setting (Guo et al., 2019; Elie et al., 2020), and the sample complexity bounds depend on the number of states and actions, which are unacceptable when the state or action spaces are very large. To the best of our knowledge, the only work studying the sample complexity of MFRL in the function approximation setting is (Pásztor et al., 2023). However, Pásztor et al. (2023) only consider MFC setting, and are limited to near-deterministic transitions with sub-Gaussian noises, which cannot model the transition distributions with multiple modes.

Motivated by these limitations, we focus on the general model-based function approximation setting, where the state-action spaces can be arbitrarily large but we have access to a model class $\mathcal{M}$ to approximate the dynamics of the Mean-Field system. The key question we would like to address is:

*What is the **fundamental** sample efficiency of Mean-Field RL with model function approximation?*

In contrast with previous literature (especially in MFG setting), by "fundamental", we aim to understand the sample efficiency *with the minimal (most basic) assumptions* including realizability (Assump. A) and Lipschitz continuity (Assump. B) without additional strong structural assumptions like contractivity or monotonicty (Guo et al., 2019; Perrin et al., 2020; Pérolat et al., 2022; Elie et al., 2020). We treat them as fundamental assumptions because realizability ensures a good approximation exists so the learning is possible, and the Lipschitz assumption, as we will show later, guarantees the existence of the learning objective in MFG setting. Moreover, we only consider the trajectory sampling model (Def. 3.3), which is much milder than the generative model assumptions (Guo et al., 2019).

To address our key question, in Sec. 4 we first propose a new notion called Mean-Field Model-Based Eluder Dimension (MF-MBED). MF-MBED, generalized from Eluder Dimension in single-agent value approximation setting (Russo and Van Roy, 2013), characterizes the complexity of mean-field function classes including but not limited to tabular setting. We also provide concrete examples of mean-field model classes with low MF-MBED, such as (generalized) linear MF-MDP, near-deterministic transition with Gaussian noise, etc. In Sec. 5, we develop sample efficient model-learning algorithms for MFRL based on Maximal Likelihood Estimation (MLE), which can achieve sample complexity only polynomial in MF-MBED for both MFC and MFG.

We highlight our main contributions in the following:

- We introduce a new notion called Mean-Field Model-Based Eluder Dimension (MF-MBED) to measure the complexity of any given Mean-Field model function class, and identify concrete examples exhibit low MF-MBED.

- For MFG setting, we propose the first MLE-based algorithm which is capable of addressing the exploration challenge while imposing minimal structural assumption. Further, we establish the first fundamental sample complexity result for function approximation setting, which only has polynomial dependence on MF-MBED, without explicit dependence on the number of states and actions.

- On the technical level, we establish MLE learning guarantees for Mean-Field setting, while previous results are limited to single-agent setting. Notably, the dependence on state density in transition function introduces unique challenges in analysis, which we overcome by establishing close connections between model class complexity, MLE error, and the MFC and MFG objectives.

## 2 RELATED WORK

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 in mean-field setting and defer the related works in single-agent or general multi-agent setting to Appx. A.1.

**Finite-horizon Non-Stationary MFG** The finite-horizon framework considered here is closely related to Lasry-Lions games (Perrin et al., 2020; Pérolat et al., 2022; Geist et al., 2022), where continuous-time dynamics were analyzed without exploration considerations under monotonicity assumptions on rewards. Most existing works consider additional structural assumptions like contractivity (Guo et al., 2019) monotonicity

(Pérolat et al., 2022). We defer to Remark 3.1 for a more detailed comparison among these assumptions. Besides, most previous literature (Elie et al., 2020) requires a planning oracle that can return a trajectory for arbitrary state density, even if it can not be induced by any policy. In contrast, our work focuses on understanding the fundamental exploration guarantees and identifying bottlenecks associated with finite-horizon MFC and MFG. Consequently, our findings extend beyond MFG/MFC scenarios that adhere to restrictive conditions.

**Stationary MFG** Alternative to the finite-horizon non-stationary MFG formulation, there exist works on the stationary MFG formulation (Anahtarci et al., 2023; Xie et al., 2021; Yardim et al., 2023; Cui and Koeppl, 2021). In this formulation, the transition and reward functions at each step are conditioned on the stationary density rather than on the evolving density across time, which poses a limitation. Furthermore, existing results typically require strong Lipschitz continuity assumptions as well as non-vanishing regularization (Anahtarci et al., 2023; Cui and Koeppl, 2021).

**Statistical Efficiency Results for MFC** In terms of statistical efficiency considerations, a related work (Pásztor et al., 2023) analyzes the MFC setting in an information gain framework. Our results capture different learnable function classes. Our low MF-MBED model classes can capture multi-modal transition distribution (e.g. the linear setting), while their algorithm and analysis are limited to near-deterministic transition with random noise (unimodal transition distribution). Besides, our framework encompasses certain special cases in (Pásztor et al., 2023), as a result of our Prop. 4.5 and the equivalence between Eluder Dimension and Information Gain in RKHS space (Jin et al., 2021). However, it is worthy noting a limitation in our approach: (Pásztor et al., 2023) can cover 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.

**Other MFG/MFC Settings** There also exists a variety of different settings in which the MFGs formalism has been utilized, for instance in linear quadratic MFG (Guo et al., 2022) and MFGs on graphs (Yang et al., 2018; Gu et al., 2021). (Angiuli et al., 2022) 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 (Carmona et al., 2023). In our work, we do not assume the observability of the population.

# 3 BACKGROUND

## 3.1 Setting and Notation

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 $\mu_1$ is the fixed initial distribution known to the learner, $\mathcal{S}$ and $\mathcal{A}$ are the state and action space. Without loss of generality, we assume the state and action spaces are discrete but can be arbitrarily large. We assume the state-action spaces 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 $r := \{r_h\}_{h=1}^H$ 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 $r_h : \mathcal{S}_h \times \mathcal{A}_h \times \Delta(S_h) \rightarrow [0, \frac{1}{H}]$. Without loss of generality, we assume that the reward function is known, but our techniques can be extended when it is unknown and a reward function class is available. We use $M^*$ to denote the true model with transition function $\mathbb{P}_{T^*}$.

In this paper, we only consider 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 $\forall h \in [H]$:

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

where:

$$\Gamma^\pi_{M,h}(\mu_h)(\cdot) := \sum_{s_h, a_h} \mu_h(s_h)\pi(a_h|s_h)\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h)$$

and we use $\mu^\pi_{M,h}$ to denote the density induced by $\pi$ under model $M$ 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 $\pi$. We will use bold font $\boldsymbol{\mu} := \{\mu_1, ..., \mu_H\}$ to denote the collection of density for all time steps. Besides, we denote $V^\pi_{M,h}(\cdot; \boldsymbol{\mu})$ to be the value functions at step $h$ if the agent deploys policy $\pi$ in model $M$ conditioning on $\boldsymbol{\mu}$, defined by:

$$V^\pi_{M,h}(s_h; \boldsymbol{\mu}) := \mathbb{E}\Big[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}, \mu_{h'})\Big|$$
$$\forall \widetilde{h} \geq h,\ a_{\widetilde{h}} \sim \pi_{\widetilde{h}},\ s_{\widetilde{h}+1} \sim \mathbb{P}_{T,\widetilde{h}}(\cdot|s_{\widetilde{h}}, a_{\widetilde{h}}, \mu_{\widetilde{h}})\Big]. \quad (2)$$

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 $\pi$ 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. $\pi$. We will omit $\boldsymbol{\mu}$ and use $J_M(\pi)$ in shorthand when $\boldsymbol{\mu} = \boldsymbol{\mu}^\pi_M$. For simplicity, in the rest of the paper, we use

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

as a shortnote of the expectation over trajectories induced by $\pi$ 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 $\mathbb{H}(P,Q) := \sqrt{1-\sum_x\sqrt{P(x)Q(x)}}$ as the Hellinger distance. In general, when $\Omega$ is countable, we have $\sqrt{2}\mathbb{H}(P,Q) \geq \mathbb{TV}(P,Q) = \frac{1}{2}\|P-Q\|_1$, where $\|\cdot\|_1$ is the $l_1$-distance.

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

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

**Mean-Field Game** In MFG, we instead want to find a Nash Equilibrium (NE) policy s.t., when all the agents follow that same policy, no agent tends to deviate from 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 $\pi^*_{\text{NE}}$ satisfying $\mathcal{E}_{\text{NE}}(\pi^*_{\text{NE}}) = 0$. In MFG, the objective is to find an approximate NE $\widehat{\pi}^*_{\text{NE}}$ such that:

$$\mathcal{E}_{\text{NE}}(\widehat{\pi}^*_{\text{NE}}) \leq \varepsilon. \tag{4}$$

### 3.2 Assumptions

In this paper, we consider the general function approximation setting, where the learner has access to a model class $\mathcal{M}$ with finite cardinality ($|\mathcal{M}| < +\infty$), 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 $M := (\mathbb{P}_T, r) \in \mathcal{M}$, we have:

$$\begin{aligned}\|\mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu_h) - \mathbb{P}_{T,h}(\cdot|s_h,a_h,\mu'_h)\|_{\mathbb{TV}}& \\ \leq L_T\|\mu_h-\mu'_h\|_{\mathbb{TV}},& \quad (5)\\ \|r_h(s_h,a_h,\mu_h) - r_h(s_h,a_h,\mu'_h)\|_{\mathbb{TV}}& \\ \leq L_r\|\mu_h-\mu'_h\|_{\mathbb{TV}}.& \quad (6)\end{aligned}$$

**Remark 3.1** (Comparison with Previous Structural Assumptions in MFG Setting). The most common structural assumptions in previous finite-horizon MFG setting include contractivity (Guo et al., 2019) and monotonicity (Perrin et al., 2020; Pérolat et al., 2022). The contractivity is stronger than ours since in general it requires good smooth conditions of $\mathbb{P}_T$ w.r.t. states, actions, state densities (Yardim et al., 2023). The monotonicity instead considers the reward structures, and it is stronger in that it assumes the transition is independent of density (i.e. $L_T = 0$), so that the dynamics of the system reduces to single-agent RL.

In Prop. 3.2 below, we show that Assump. B implies the existence of a Nash Equilibrium. A similar existence result has been established in previous literature (Saldi et al., 2018) under the same conditions as our Prop. E.8. Our contribution here is a different proof based on the conjugate function and non-expansiveness of the proximal point operator. Moreover, (Saldi et al., 2018) studied infinite-horizon MDP with discounted reward, which is different from our setting.

**Proposition 3.2** (Existence of NE in MFG; Informal Version of Prop. E.8). *For every MF-MDP with discrete $\mathcal{S}$ and $\mathcal{A}$, satisfying Assump. B, there exists at least one NE policy.*

Besides, we formalize the trajectory sampling model in the following. Our trajectory sampling model is much weaker than the assumption in (Guo et al., 2019; Elie et al., 2020), which requires a planning oracle that can return a trajectory conditioning on arbitrary (even unachievable) density.

**Definition 3.3** (Trajectory Sampling Model). We assume the environment consists of an extremely large number of agents and a central controller (our algorithm/learner), and there is a representative agent `Agent`, whose observation we can receive. The central controller can compute an arbitrary policy tuple ($\widetilde{\pi}$, $\pi$) (here $\pi$ and $\widetilde{\pi}$ are not necessarily the same), distribute $\widetilde{\pi}$ to `Agent` but $\pi$ to the others, and receive the trajectory of `Agent` following $\widetilde{\pi}$ under $\mathbb{P}_{T^*,h}(\cdot|\cdot,\cdot,\mu^\pi_h)$ and $\mathbb{P}_{r,h}(\cdot|\cdot,\cdot,\mu^\pi_h)$.

## 4 MEAN-FIELD MODEL-BASED ELUDER DIMENSION

In order to avoid explicit dependence on the number of states and actions, we focus on the intrinsic complexity of the function class instead of the complexity of the state and action spaces. In this section, we introduce the notion of Mean-Field Model-Based Eluder Dimension (MF-MBED), which characterizes the complexity of an arbitrary model function class $\mathcal{M}$ via the length of the longest state-action- state density sequence $\{(s^i,a^i,\mu^i)\}_{i\in[n]}$ such that each $(s^i,a^i,\mu^i)$ introduces new "information" of $\mathcal{M}$ given the information revealed by previous data $\{(s^t,a^t,\mu^t)\}_{t=1,\ldots,i-1}$. In the following, we introduce the formal definition.

**Definition 4.1** ($\alpha$-weakly-$\varepsilon$-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 $\alpha$-weakly-$\varepsilon$-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 4.2** (The longest $\alpha$-weakly-$\varepsilon$-independent sequence)**.** We use $\text{dimE}_\alpha(\mathcal{F}, \mathbf{D}, \varepsilon)$ to denote the longest sequence $\{x_i\}_{i=1}^n \in \mathcal{X}$, such that for some $\varepsilon' \geq \varepsilon$, $x_i$ is $\alpha$-weakly-$\varepsilon'$-independent of $\{x_1, ..., x_{i-1}\}$ for all $i \in [n]$ w.r.t. $\mathcal{F}$ and $\mathbf{D}$.

**Definition 4.3** (Model-Based Eluder-Dimension in Mean-Field RL)**.** Given a model class $\mathcal{M}$, $\alpha \geq 1$ and $\varepsilon > 0$, the Model-Based Eluder Dimension in MFRL (abbr. MF-MBED) of $\mathcal{M}$ is defined as:

$$\text{dimE}_\alpha(\mathcal{M}, \varepsilon) := \max_{h\in[H]} \min_{\mathbf{D}\in\{\mathbb{TV},\mathbb{H}\}} \text{dimE}_\alpha(\mathcal{M}_h, \mathbf{D}, \varepsilon). \tag{7}$$

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 slight abuse of notation, $\mathcal{M}$ (or $\mathcal{M}_h$) here refers to the collection of transition functions of models in $\mathcal{M}$. The main difference comparing with value function approximation setting (Russo and Van Roy, 2013; Jin et al., 2021) is that, because the output of model functions are distributions instead of scalar, we use distance measure to compute the model prediction difference. Furthermore, we use $\alpha\varepsilon$ instead of $\varepsilon$ in the threshold, 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 $\text{dimE}_{\alpha_1}(\mathcal{F}, \varepsilon) \leq \text{dimE}_{\alpha_2}(\mathcal{F}, \varepsilon)$ for $\alpha_1 \geq \alpha_2$, because any $\alpha_1$-weakly-$\varepsilon$-independent sequence must be $\alpha_2$-weakly-$\varepsilon$-independent.

**Comparison with Previous Work Regarding Eluder Dimension** Multiple previous work considers using independent sequences to characterize the complexity of function classes, but most of them focus on value function approximation in the single-agent setting (Russo and Van Roy, 2013; Jin et al., 2021). To our knowledge, only limited work (Osband and Van Roy, 2014; Levy et al., 2022) has studied Eluder Dimension for model-based function approximation even in single-agent setting. (Osband and Van Roy, 2014) additionally assumed given two transition distributions in the function class, the difference between their induced future value function is Lipschitz continuous w.r.t. their mean difference, which is restrictive. More recently, (Levy et al., 2022) 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.

### 4.1 Concrete Examples with Low MF-MBED

Next, we introduce some concrete examples with low MF-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 (Jin et al., 2020).

**Proposition 4.4** (Low-Rank MF-MDP with Known Representation; Informal Version of Prop. B.4)**.** *Given a feature $\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \to \mathbb{R}^d$ and a function class $\Psi$, 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 $\alpha \geq 1$.*

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. We defer to appendix for more discussions about our technical novelty here. Basically, since the output of the model function is a probability distribution rather than a scalar, we utilize data-dependent sign functions to establish the connections between the TV-distance of distribution predictions and the error of next state features.

The second example considers deterministic transition with random noise, in order to accommodate the function class in (Pásztor et al., 2023) (see a detailed comparison in Sec. 2). 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 MF-MBED framework.

**Proposition 4.5.** *[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 $\Sigma :=$ $\text{Diag}(\sigma, ..., \sigma)$. 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 (Russo and Van Roy, 2013; Osband and Van Roy, 2014).*

**Remark 4.6.** In the Gaussian example above, the state space has to be continuous since the Gaussian distribution is defined in continuous space. Although in this paper we only consider the discrete state space, our results are based on $\mathbb{TV}$-distance and similar results can be established when the state space is continuous.

# 5 LEARNING IN MEAN-FIELD RL: AN MLE APPROACH

In this section, we develop a general Maximum Likelihood Estimation (MLE)-based learning framework for both MFC and MFG and show that given model classes with low MF-MBED, these MFRL problems are indeed tractable in sample complexity.

## 5.1 Main Algorithm

Since the MLE-based model learning for MFC and MFG share some similarities, we unify them in one algorithm and use MFC and MFG to distinguish algorithm steps for these two different objectives. Our main algorithm is presented in Alg. 1, where we omit the usage of rewards in learning to avoid redundancy in analysis.

The algorithm includes two parts: policy selection (Line 8-16) 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 $\pi^{k+1}$ 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. 2, 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 $\pi^{k+1}$ 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 $\pi^{k+1}$, we also collect trajectories sampled by policy $\widetilde{\pi}^{k+1}$ conditioning on the density induced by $\pi^{k+1}$. As we will explain in Lem. 6.5, those additional samples are necessary to control the estimation error of exploitability. Finally, we return the policy with the minimal optimistic exploitability among $\{\pi^{k+1}\}_{k=1}^K$.

## 5.2 Main Results on Statistical Efficiency

We state our main results below, which establish the sample complexity of learning MFC/MFG in Alg. 1. The formal version (Thm. D.5 and Thm. D.6) and the proofs are deferred to Appx. D.

**Algorithm 1:** A General Maximal Likelihood Estimation-Based Algorithm for Mean-Field RL

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

**Theorem 5.1** (Main Results (Informal))**.** *Under Assump.A, B, by running Alg. 1 with* [1]

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

*with* $\varepsilon_0 = O(\frac{L_T \varepsilon}{\alpha H(1+L_r H)(1+L_T H)((1+L_T)^H - 1)})$,

[1] We omit log-dependence on $\varepsilon, \delta, \text{dimE}, |\mathcal{M}|, H$ and Lipschitz factors in $\widetilde{O}$, and it's same for Thm. 5.2

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

As stated in our main results, both MFC and MFG can be sample efficiently solved as long as the model function classes has low MF-MBED. Also note that our sample complexity bounds are universal since MF-MBED can be defined for arbitrary mean-field model function classes. As reflected by Thm. 5.1, comparing with previous literature requiring additional structural assumptions (Guo et al., 2019; Xie et al., 2021; Perrin et al., 2020; Pérolat et al., 2022; Elie et al., 2020), our algorithm can tackle the exploration challenges in more general setting where only realizability and Lipschitz continuity assumptions are satisfied.

As for the dependence of Lipschitz factors, similar exponential dependence of $L_T$ has been reported in previous literature (Pásztor et al., 2023). We conjecture that such exponential dependence reflects some fundamental difficulties if we only make minimal assumptions. In the following, we show how to avoid exponential terms under additional contraction assumptions.

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

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

where $\Gamma^\pi_{M,h}(\mu_h)$ is defined in Eq. (1). According to (Yardim et al., 2023), Assump. C 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).

We summarize the main results given Assump. C below, where the formal versions are also included in Thm. D.5 and Thm. D.6. Comparing with Thm. 5.1, the sample complexity exhibits only polynomial dependence on Lipschitz factors.

**Theorem 5.2** (Main Results (Informal))**.** *Under Assump.A, B and C, by choosing:*

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

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

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

### 5.3 Generalization to Infinite Model Class and Continuous Setting

When the model class $\mathcal{M}$ contains infinite model candidates, our results can be generalized by the following steps. Firstly, we can find an $\varepsilon$-cover $\mathcal{M}_\varepsilon$ w.r.t. the distance $d(\cdot,\cdot)$ defined by:

$$d(M,M'):=\max_{\pi,\mu}\max\{\mathbb{E}_{\pi,M|\mu}[\sum_h\log\frac{\mathbb{P}_{M,h}(s'_h|s_h,a_h,\mu_h)}{\mathbb{P}_{M',h}(s'_h|s_h,a_h,\mu_h)}],\ \mathbb{E}_{\pi,M'|\mu}[\sum_h\log\frac{\mathbb{P}_{M',h}(s'_h|s_h,a_h,\mu_h)}{\mathbb{P}_{M,h}(s'_h|s_h,a_h,\mu_h)}]\}$$

which aligns with our MLE method. For $\mathcal{M}_\varepsilon$, Assump. A may not hold because of the discretization, but there exists a model $\widehat{M}^*\in\mathcal{M}_\varepsilon$ close to $M^*$ under distance $d$ by definition. Then, all we need to do is to revise line 10 of Alg. 1 to

$$\widehat{\mathcal{M}}^k\leftarrow\{M\in\mathcal{M}|l^k_{MLE}(M)\geq\max_{M\in\mathcal{M}}l^k_{MLE}(M)-\log\frac{2|\mathcal{M}|KH}{\delta}-O(\varepsilon)\},$$

where we have additional tolerance at level $O(\varepsilon)$. Then we can extend our sample complexity results and it will only depend on $\log|\mathcal{M}_\varepsilon|$ instead. Besides, assuming low log covering number is common in previous literature, e.g. Def. 13 in (Jiang et al., 2017), Def. 3 in (Jin et al., 2021), etc.

Our algorithm, analysis, and the notion of "Model-Based Eluder Dimension" can be extended to the case with continuous compact $\mathcal{S},\mathcal{A}$ spaces with minor modifications (such as replacing $\sum_{s,a}$ with $\int dsda$ and $l_1$ distance with total-variation distance). The only exception is Prop. 3.2, which relies on fixed point theorem in finite space, but it is reasonable to assume the existence of Nash equilibrium when the state and action spaces are continuous.

## 6 PROOF SKETCH

In this section, we provide proof sketch to establish Thm. 5.1. Intuitively, the proofs consist of two parts. Firstly, we provide an upper bound for the accumulative model prediction error by the MF-MBED, which we further connect with our learning objective in the second step.

**Step 1: Upper Bound Model Prediction Error with MF-MBED** First of all, in Thm. 6.1 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 6.1.** *[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 $h \in [H]$:*

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

*Besides, for MFG branch, we additionally have:*

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

The key difficulty in Mean-Field setting is the dependence of density in transition function. Since we do not know $\mu_{M^*,h}^\pi$, in Line 9 in Alg. 1, we compute the likelihood conditioning on $\mu_{M,h}^\pi$, which is accessible for each $M$. Therefore, in Thm. 6.1, we can only guarantee $M$ aligns with $M^*$ conditioning on their own density $\boldsymbol{\mu}_M^\pi$ and $\boldsymbol{\mu}_{M^*}^\pi$, respectively. However, to ensure low MF-MBED can indeed capture important practical models, the MF-MBED in Def. 4.3 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. 6.2 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 6.2.** *[Model Difference Conversion] 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 $\pi$, under Assump. B, we have:*

$$\mathbb{E}_{\pi,M}[\sum_{h=1}^{H} \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_{M,h}^\pi) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu_{M,h}^\pi)\|_{\mathbb{TV}}] \\ \leq (1 + L_T H)\mathbb{E}_{\pi,M}[\sum_{h=1}^{H} \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_{M,h}^\pi) \\ - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu_{\widetilde{M},h}^\pi)\|_{\mathbb{TV}}], \tag{8}$$

*and*

$$\mathbb{E}_{\pi,M}[\sum_{h=1}^{H} \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_{M,h}^\pi) - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu_{\widetilde{M},h}^\pi)\|_{\mathbb{TV}}] \\ \leq \mathbb{E}_{\pi,M}[\sum_{h=1}^{H} (1 + L_T)^{H-h} \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_{M,h}^\pi) \\ - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu_{M,h}^\pi)\|_{\mathbb{TV}}]. \tag{9}$$

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 MF-MBED. We defer the proof to Appx. B.3.

**Lemma 6.3.** *Under the condition as Def. 4.1, 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 6.4.** *[Value Difference Lemma for MFC] Given an arbitrary model $M$ with transition function $\mathbb{P}_T$, and an arbitrary policy $\pi$, under Assump. B, we have:*

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

By Thm. 6.1 and Eq. (8) in Thm. 6.2, 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 $\mu_{M^*}^{\pi^1}, ..., \mu_{M^*}^{\pi^k}$. On good concentration events, the condition for Lem. 6.3 is satisfied, and as a result of Thm. 6.4 and Eq. (9), 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. 2, we can establish the sample complexity guarantee in Thm. 5.1.

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:

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

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

**Lemma 6.5.** *[Value Difference Lemma for MFG] Given two arbitrary model $M$ and $\widetilde{M}$, and two policies $\pi$ and $\widetilde{\pi}$, we have:*

$$|\Delta_M(\widetilde{\pi}, \pi) - \Delta_{\widetilde{M}}(\widetilde{\pi}, \pi)| \\ \leq \mathbb{E}_{\widetilde{\pi}, M|\boldsymbol{\mu}_M^\pi}[\sum_{h=1}^{H} \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_{M,h}^\pi)$$

$$- \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h})\|_{\mathbb{TV}}]$$
$$+ (2L_r H + 1)\mathbb{E}_{\pi,M}[\sum_{h=1}^{H} \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu^{\pi}_{M,h})$$
$$- \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu^{\pi}_{\widetilde{M},h})\|_{\mathbb{TV}}]. \quad (10)$$

As we can see, to control the exploitability, we require the model can predict well on the data distribution induced by both $\pi^{k+1}$ and $\widetilde{\pi}^{k+1}$ conditioning on $\boldsymbol{\mu}^{\pi^{k+1}}_{M^*}$, which motivates our formulation of Def. 3.3. By combining Lem. 6.5 and theorems in the first part, we finish the proof.

**A Different Model Difference Conversion under Assump. C** In the following theorem, we provide a model difference conversion conditioning on different densities under Assump. C to replace Eq. (9) in Thm. 6.2, which is the key observation to avoid exponential dependence.

**Theorem 6.6.** *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* $\pi$*, under Assump. B and Assump. C, we have:*

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

# 7 DISCUSSION AND OPEN PROBLEMS

In this paper, we study the statistical efficiency of function approximation in MFRL. We propose the notion of MF-MBED and an MLE based algorithm, which can guarantee to efficiently solve MFC and MFG given function classes satisfying basic assumptions including realizablility and Lipschitz continuity. Under additional structural assumptions, the exponential dependence on Lipschitz factors in sample complexity bounds could be avoided. In the following, we discuss some potentially interesting open problems.

**Tighter Sample Complexity Upper Bounds and Lower Bounds** Although in this paper, we show that both MFC and MFG can be solved under the same MLE-based model learning framework and establish the same sample complexity depending on MF-MBED, it does not implies MFC and MFG fundamentally have the same sample efficiency. It would be an interesting direction to investigate other complexity measure which may provide tighter complexity upper bounds than our MF-MBED. Besides the sample complexity upper bound, another interesting question would be the sample complexity lower bound for MFC and MFG, and whether there exists separation between MFC and MFG in terms of sample efficiency.

**Computational Efficiency** Since we focus on the fundamental sample efficiency, we ignore complicated computation processes and abstract them as computational oracles. For MFC setting, similar maximization oracles has been treated as mild assumptions . For MFG setting, although previous literature (Guo et al., 2019; Pérolat et al., 2022) has provided concrete implementations given additional structural assumptions like contractivity and monotonicity, whether NE can be solved efficiently under more general setting is an open problem. Therefore, another important direction is to combine our results with optimization techniques to design computationally efficient algorithms.

**Model-Free Methods in Function Approximation Setting** In this paper, we consider the model-based methods. Given the popularity of model-free methods in single-agent setting, it remains an open problem what the sample efficiency of model-free methods with general function approximations are. Even though, we prefer model-based function approximation in mean-field setting because model-free methods suffer from some additional challenges seems intractable. For pure value-based methods, without explicit model estimation, one may not infer the density, and therefore may not estimate the true value function accurately. Policy-based methods would be more promising and has been applied in tabular setting (Guo et al., 2019; Yardim et al., 2023), where by introducing the policy, one just need to focus on the estimation of values conditioning on policy and density estimation will be unimportant. But when generalizing to function approximation setting, the value function class should be powerful enough to approximate the value functions regarding all the policies occurred in the learning process, which might be very strong assumptions. We leave the investigation of this direction to the future.

### Acknowledgements

The work is supported by ETH research grant, NCCR Automation, Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343. and SNSF Starting Grant. The authors would like to thank Andreas Krause for valuable discussion and thank Lars Lorch for valuable feedbacks on paper writing.

## References

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

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

Angiuli, A., Fouque, J.-P., and Lauriere, M. (2021). Reinforcement learning for mean field games, with applications to economics. *arXiv preprint arXiv:2106.13755*.

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

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

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

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

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

Bensoussan, A., Frehse, J., Yam, P., et al. (2013). *Mean field games and mean field type control theory*, volume 101. Springer.

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

Carmona, R., Laurière, M., and Tan, Z. (2023). Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *The Annals of Applied Probability*.

Chen, Z., Li, C. J., Yuan, H., Gu, Q., and Jordan, M. (2022a). A general framework for sample-efficient function approximation in reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

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

Cousin, A., Crépey, S., Guéant, O., Hobson, D., Jeanblanc, M., Lasry, J.-M., Laurent, J.-P., Lions, P.-L., Tankov, P., Guéant, O., et al. (2011). Mean field games and applications. *Paris-Princeton lectures on mathematical finance 2010*, pages 205–266.

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

Cui, Q., Zhang, K., and Du, S. (2023). Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2651–2652. PMLR.

De Paola, A., Trovato, V., Angeli, D., and Strbac, G. (2019). 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.

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

Elie, R., Perolat, J., Laurière, M., Geist, M., and Pietquin, O. (2020). 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.

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

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. (2022). Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 489–497.

Gu, H., Guo, X., Wei, X., and Xu, R. (2021). Mean-field multi-agent reinforcement learning: A decentralized network approach. *arXiv preprint arXiv:2108.02731*.

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

Guo, X., Xu, R., and Zariphopoulou, T. (2022). Entropy regularization for mean field games with learning. *Mathematics of Operations Research*.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. (2021). Towards general function approximation in zero-sum markov games. In *International Conference on Learning Representations*.

Huang, J., Chen, J., Zhao, L., Qin, T., Jiang, N., and Liu, T.-Y. (2022). Towards deployment-efficient reinforcement learning: Lower bound and optimality. In *International Conference on Learning Representations*.

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

Ismail, Z. H., Sariff, N., and Hurtado, E. (2018). A survey and analysis of cooperative multi-agent robot systems: challenges and directions. *Applications of Mobile Robots*, pages 8–14.

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

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

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

Jin, C., Liu, Q., Wang, Y., and Yu, T. (2023). V-learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning. *Mathematics of Operations Research*.

Jin, C., Liu, Q., and Yu, T. (2022). The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR.

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

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

Lee, J. W., Park, J., Jangmin, O., Lee, J., and Hong, E. (2007). 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.

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

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. (2022a). When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR.

Liu, Q., Szepesvári, C., and Jin, C. (2022b). Sample-efficient reinforcement learning of partially observable markov games. *Advances in Neural Information Processing Systems*, 35:18296–18308.

Mahajan, A. (2021). Reinforcement learning in stationary mean-field games.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. (2024). Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 25(6):1–76.

Modi, A., Jiang, N., Tewari, A., and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR.

Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

Ni, C., Song, Y., Zhang, X., Ding, Z., Jin, C., and Wang, M. (2023). Representation learning for low-rank general-sum markov games. In *The Eleventh International Conference on Learning Representations*.

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

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and trends in Optimization*, 1(3):127–239.

Pásztor, B., Krause, A., and Bogunovic, I. (2023). Efficient model-based multi-agent mean-field reinforcement learning. *Transactions on Machine Learning Research*.

Pérolat, J., Perrin, S., Elie, R., Laurière, M., Piliouras, G., Geist, M., Tuyls, K., and Pietquin, O. (2022). Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1028–1037.

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

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

Saldi, N., Basar, T., and Raginsky, M. (2018). Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Shalev-Shwartz, S. and Singer, Y. (2007). *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). 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.

Uehara, M., Zhang, X., and Sun, W. (2021). Representation learning for online and offline rl in low-rank mdps. In *International Conference on Learning Representations*.

Wang, Y., Liu, Q., Bai, Y., and Jin, C. (2023). Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. In *Proceedings of Thirty Sixth Conference on Learning Theory*, Proceedings of Machine Learning Research. PMLR.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. (2020). Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR.

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

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. (2022). The role of coverage in online reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

Xiong, W., Zhong, H., Shi, C., Shen, C., and Zhang, T. (2022). A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning*, pages 24496–24523. PMLR.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018). Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR.

Yardim, B., Cayci, S., Geist, M., and He, N. (2023). Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pages 39722–39754. PMLR.

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

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

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

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

## Contents

# A Extended Introduction

## A.1 Additional Related Works

**Single-Agent RL with General Function Approximation** Recently, beyond tabular RL (Auer et al., 2008; Azar et al., 2017; Jin et al., 2018), there are significant progress on sample complexity analysis in linear function approximation setting (Jin et al., 2020; Zanette et al., 2020; Agarwal et al., 2020; Modi et al., 2024; Uehara et al., 2021; Huang et al., 2022) or more general function approximation settings (Russo and Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Jin et al., 2021; Du et al., 2021; Xie et al., 2022; Foster et al., 2021; Chen et al., 2022a; Ayoub et al., 2020). However, the MFRL setting is significantly different from single-agent RL because of the dependence on the state density in transition and reward functions. The function complexity measure, especially for value-based function class, and the corresponding algorithms in single-agent RL cannot be trivially generalized to MFRL.

**Multi-Agent RL** Sample complexity of learning in Markov Games has been studied extensively in a recent surge of works (Jin et al., 2023; Bai et al., 2020; Chen et al., 2022b; Zhang et al., 2019, 2021; Xiong et al., 2022). A few recent works also consider learning Markov Games with linear or general function approximation (Xie et al., 2020; Huang et al., 2021; Jin et al., 2022; Ni et al., 2023). None of these results can be directly extended to Mean-Field RL.

Recently, (Wang et al., 2023; Cui et al., 2023) 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 agents can be largely different, there are still some restrictions when generalizing their results 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, the sample complexity of their algorithms depend on the number of agents (although polynomially instead of exponentially), which implies that their algorithms 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 (Liu et al., 2022a) or Partial Observable Markov Games (Liu et al., 2022b).

# 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 (Russo and Van Roy, 2013).

**Definition B.1** ($\varepsilon$-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 $\varepsilon$-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 $\overline{\dim\mathrm{E}}(\mathcal{F}, \varepsilon)$ 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 $\varepsilon$-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. 6.1, 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. 4.5, 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 $\dim\mathrm{E}_\alpha(\mathcal{M}, \mathbb{TV}, \varepsilon)$ 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 $\|\sum_{s'\in\mathcal{S}} \phi(s, a, \mu, s')g(s')\|_2 \le C_\phi$*[2]

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

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

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

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

where we define:

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

In the following, for simplicity, we use

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

as a short note. Also note that,

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

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 $\varepsilon$-independent of $\{x_1, ..., x_{i-1}\}$ for all $i \in [n]$. Then, by definition, for each $i$, there exists $\theta_1^i, \theta_2^i$ such that:

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

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,

$$\begin{aligned} \|\theta_1^i - \theta_2^i\|_{\Lambda^i}^2 =& \lambda \|\theta_1^i - \theta_2^i\|^2 + \sum_{t=1}^{i-1} \Big((\theta_1^i - \theta_2^i)^\top v_{\theta_1^t,\theta_2^t}(s^t, a^t, \mu^t)\Big)^2 \\ \le& 4\lambda C_\theta^2 + \sum_{t=1}^{i-1} \Big((\theta_1^i - \theta_2^i) \sum_{s'} \Phi(s^t, a^t, \mu^t) \psi(s') g_{\theta_1^t,\theta_2^t}(s^t, a^t, \mu^t, s')\Big)^2 \end{aligned}$$

[2] Similar normalization assumptions is common in previous literatures (Agarwal et al., 2020; Modi et al., 2020; Uehara et al., 2021)

$$
\begin{aligned}
&=4\lambda C_\theta^2 + 4\sum_{t=1}^{i-1} \|\mathbb{P}_{\theta_1^t}(\cdot|s^t,a^t,\mu^t), \mathbb{P}_{\theta_2^t}(\cdot|s^t,a^t,\mu^t)\|_{\mathbb{TV}} && (|g_{\theta_1^t,\theta_2^t}(\cdot,\cdot,\cdot,\cdot)|=1)\\
&\leq 4\lambda C_\theta^2 + 4\varepsilon^2.
\end{aligned}
$$

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,

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

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

**Linear Combined Model with State-Action-Dependent Weights** In (Modi et al., 2020), 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 MF-MBED.

### B.2.2 Example 2: Linear MDP with Known Feature

**Proposition B.4** (Low-Rank MF-MDP; Formal Version of Prop. 4.4)**.** *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 $\Psi$ satisfying $\forall \psi \in \Psi,\ \forall s' \in \mathcal{S},\ \forall g : \mathcal{S} \rightarrow \{-1,1\},\ \|\sum_{s'} \psi(s')g(s')\|_2 \leq C_\Psi$, consider the following model class:*

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

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

*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 $\varepsilon$-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

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

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 $v_{\psi,\widetilde{\psi}}(s,a,\mu) := \sum_{s'}(\psi(s') - \widetilde{\psi}(s'))g_{\psi,\widetilde{\psi}}(s,a,\mu,s')$ as a shortnote. Therefore, for each $i$,

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

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,

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

Therefore,

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

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 (Du et al., 2021):

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

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

In the next theorem, we show that, the MF-MBED of kernel MF-MDP generalized from kernel MDP in single-agent setting (Jin et al., 2021) can be upper bounded by the effective 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 $g : \mathcal{S} \rightarrow \{-1, 1\}$, $\|\sum_{s'\in\mathcal{S}} \psi(s')g(s')\|_{\mathcal{H}} \leq 1$* [3]. *Consider the model class $\mathcal{P}_{\Psi,\mathcal{H}}$ defined by:*

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

*For $\alpha \geq 1$, we have*

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

*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 $\varepsilon$-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:

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

and

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

where we define:

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

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 (Russo and Van Roy, 2013) to MF-MDP.

[3] To align with (Jin et al., 2021), we assume $\psi$ is normalized.

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

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

*For any $\alpha \ge 1$, 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 $\varepsilon$-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:

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

Besides,

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

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

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

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

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

By a similar discussion, we have:

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

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 5: Deterministic Transition with Gaussian Noise

**Proposition 4.5.** *[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 $\Sigma := \mathrm{Diag}(\sigma,...,\sigma)$. 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 (Russo and Van Roy, 2013; Osband and Van Roy, 2014).*

*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 $\alpha$-weakly-$\varepsilon$-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

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

and

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

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,

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

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 $(s^i,a^i,\mu^i,t_i)$ 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,

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

□

### B.3 From Eluder Dimension to Regret Bound

**Lemma B.8.** *Under the condition and notation as Def. 4.1, 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:*

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

*Proof.* We first show that, for some $k$, if $\mathbf{D}(f_k, f^*)(x_k) > \alpha\varepsilon$, then $x_k$ is $\varepsilon$-dependent on at most $\beta/\varepsilon^2$ disjoint sub-sequence in $\{x_1, ..., x_{k-1}\}$. To see this, by Def. 4.3, if $\mathbf{D}(f_k, f^*)(x_k) > \alpha\varepsilon$ and $x_k$ is $\alpha$-weakly-$\varepsilon$-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 $\beta/\varepsilon^2$.

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 $\alpha$-weakly-$\varepsilon$-dependent on $L := \lfloor \kappa / \mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon) \rfloor$ 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 $\alpha$-weakly-$\varepsilon$-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 $\alpha$-weakly-$\varepsilon$-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 MF-MBED is bounded, $B_l$ can not be larger than $\mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon)$ if the above process continues. Therefore, the process must stop before $j \leq L \cdot \mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon) \leq \kappa$.

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 $\alpha$-weakly-$\varepsilon$-dependent with at most $\beta/\varepsilon^2$ disjoint sub-sequence of $\{x_{k_1}, ... x_{k_{j-1}}\}$, and on the other hand, $x_{k_j}$ is $\alpha$-weakly-$\varepsilon$-dependent on at least $L := \lfloor \kappa / \mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon) \rfloor$ disjoint sub-sequence of $\{x_{k_1}, ... x_{k_{j-1}}\}$. Therefore, we have:

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

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

**Lemma 6.3.** *Under the condition as Def. 4.1, 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 $\sum_{k=1}^K \boldsymbol{D}(f_k, f^*)(x_k) = O(\sqrt{\beta K \, dimE_\alpha(\mathcal{M}, \varepsilon)} + \alpha K \varepsilon)$.*

*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:

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

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

$$\begin{aligned} \sum_{k=1}^K e_k \leq & \alpha K \varepsilon + \sum_{k=1}^K \mathbb{I}[e_k > \alpha\varepsilon] e_k \\ & \leq \alpha K \varepsilon + (\mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon) + 1) C + \sum_{k = \mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon) + 2}^K \sqrt{\frac{\beta \mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon)}{t - \mathrm{dimE}_\alpha(\mathcal{F}, \varepsilon)}} \end{aligned}$$

(Recall the constant $C$ is the upper bound for $\mathbf{D}(f, f^*)(x)$)

$$\le \alpha K\varepsilon + (\mathrm{dimE}_\alpha(\mathcal{F},\varepsilon)+1)C + \sqrt{\beta \mathrm{dimE}_\alpha(\mathcal{F},\varepsilon)} \sum_{t=\mathrm{dimE}_\alpha(\mathcal{F},\varepsilon)+1}^{K} \frac{1}{\sqrt{t-\mathrm{dimE}_\alpha(\mathcal{F},\varepsilon)}}dt$$
$$=O(\sqrt{\beta K \mathrm{dimE}_\alpha(\mathcal{F},\varepsilon)} + \alpha K \varepsilon).$$

□

# 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 $\delta\in(0,1)$:*

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

*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:

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

Here the last but two step is because:

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

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

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 $\pi^i$ and $\widetilde{\pi}^i$.

**Lemma C.2.** *Let $l : \Pi \times \Pi \times (\mathcal{S}\times\mathcal{A}\times\mathcal{S})^H \times (\mathcal{S}\times\mathcal{A}\times\mathcal{S})^H \to \mathbb{R}$ 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,

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

□

**Theorem 6.1.** *[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 $h \in [H]$:*

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

*Besides, for MFG branch, we additionally have:*

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

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

$$l_M(\pi,\widetilde{\pi},\mathcal{Z}) := \begin{cases} \frac{1}{2}\log\frac{f_M^{\pi,\widetilde{\pi}}(\mathcal{Z})}{f_{M^*}^{\pi,\widetilde{\pi}}(\mathcal{Z})}, & \text{if } f_{M^*}^{\pi,\widetilde{\pi}}(\mathcal{Z}) \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

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

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

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 $M \in \widehat{\mathcal{M}}^k$:

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

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

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

For any $i \in [k]$ and for arbitrary random variable $s_h^i, a_h^i \in \mathcal{Z}_{cond}^i$ and $s_{h+1}'^i \in \mathcal{Z}_{pred}^i$, we have:

$$\begin{aligned} & \mathbb{E}_{\mathcal{Z}_{cond}^i}[1 - \sum_{\mathcal{Z}_{pred}^i}\sqrt{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i) f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}|\pi^i,\widetilde{\pi}^i,M^*] \\ =& \mathbb{E}_{\mathcal{Z}_{cond}^i}[1 - \sum_{s_{h+1}'^i}\sqrt{\mathbb{P}_{T,h}(s_{h+1}'^i|s_h^i,a_h^i,\mu_{M,h}^{\pi^i})\mathbb{P}_{T^*,h}(s_{h+1}'^i|s_h^i,a_h^i,\mu_{M^*,h}^{\pi^i})} \sum_{\mathcal{Z}_{pred}^i\setminus\{s_{h+1}'^i\}}\sqrt{f_M^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i) f_{M^*}^{\pi^i,\widetilde{\pi}^i}(\mathcal{Z}^i)}|\pi^i,\widetilde{\pi}^i,M^*] \\ & \qquad (\text{Independence between } s_{h+1}'^i \text{ and } \mathcal{Z}^i\setminus\{s_{h+1}'^i\} \text{ conditioning on } \mathcal{Z}_{cond}^i) \end{aligned}$$

$$\begin{aligned}
\geq & \mathbb{E}_{\mathcal{Z}^i_{cond}}[1 - \sum_{s'^i_{h+1}} \sqrt{\mathbb{P}_{T,h}(s'^i_{h+1}|s^i_h, a^i_h, \mu^{\pi^i}_{M,h})\mathbb{P}_{T^*,h}(s'^i_{h+1}|s^i_h, a^i_h, \mu^{\pi^i}_{M^*,h})}|\pi^i, \widetilde{\pi}^i, M^*] \qquad (\sqrt{ab} \leq \tfrac{a+b}{2}) \\
= & \mathbb{E}_{s^i_h, a^i_h}[1 - \sum_{s'^i_{h+1}} \sqrt{\mathbb{P}_{T,h}(s'^i_{h+1}|s^i_h, a^i_h, \mu^{\pi^i}_{M,h})\mathbb{P}_{T^*,h}(s'^i_{h+1}|s^i_h, a^i_h, \mu^{\pi^i}_{M^*,h})}|\pi^i, \widetilde{\pi}^i, M^*] \\
= & \mathbb{E}_{\pi^i, M^*}[\mathbb{H}^2(\mathbb{P}_{T,h}(\cdot|s^i_h, a^i_h, \mu^{\pi^i}_{M,h}),\ \mathbb{P}_{T^*,h}(\cdot|s^i_h, a^i_h, \mu^{\pi^i}_{M^*,h}))].
\end{aligned}$$

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:

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

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. 6.1, for any $k \in [K]$, $M \in \widehat{\mathcal{M}}^k$, and a fixed $h \in [H]$, we have:*

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

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

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

(Lem. D.1; Cauchy-Schwarz inequality;)

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:

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

Similarly, we have:

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

By similar discussion, we have:

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

□

**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:*

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

*Proof.* We first take a look at the data $(\widetilde{s}^k_h, \widetilde{a}^k_h, \widetilde{s}'^k_{h+1})$ collected by $(\widetilde{\pi}^i, \pi^i)$ and the Eluder Dimension w.r.t. the Hellinger distance. On the event in Thm. 6.1 (which implies Corollary C.3) and Lem. D.4, there exists an absolute constant $c_{\mathbb{TV}}$, 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:

$$\sum_{i=1}^k \mathbb{TV}^2(\mathbb{P}_{T^*,h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M^*,h}), \mathbb{P}_{\widehat{T}^{k+1},h}(\cdot|\widetilde{s}^i_h,\widetilde{a}^i_h,\mu^{\pi^i}_{M^*,h})) \le c_{\mathbb{TV}}(1+L_T^2H^2)\log\frac{2|\mathcal{M}|KH}{\delta}. \tag{11}$$

By Lem. 6.3, there exists some constant $c'_{\mathbb{TV}}$, for any $\varepsilon_0$, we have:

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

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

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

Combine them together, for some constant $c$, we have:

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

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

Then, we can conduct similar discussion for the data $(s^i_h, a^i_h, s^i_{h+1})$ collected by $(\pi^{k+1}, \pi^{k+1})$, and for some constant $c'$, we have:

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

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

**Algorithm 2:** 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 $\pi$, we have:*

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

*Besides, under Assump. B, we have:*

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

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

**Proof for Eq.** (12)

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

For the first part, we have:

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

For the second part, we have:

$$\sup_{\bar{\mathcal{S}}'\subset\mathcal{S}} |\sum_{s_{h+1}\in\bar{\mathcal{S}}'} \sum_{s_h,a_h} \mu^\pi_{M,h}(s_h)\pi(a_h|s_h)(\mathbb{P}_{T,h}(s_{h+1}|s_h,a_h,\mu^\pi_{M,h}) - \mathbb{P}_{\widetilde{T},h}(s_{h+1}|s_h,a_h,\mu^\pi_{\widetilde{M},h}))|$$

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

Therefore,

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

**Proof for Eq.** (13) Starting with the first inequality of Eq. (14) and applying the Assump. B, we directly have:

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

□

**Lemma D.2.** *Under Assump. B and Assump. C, we have:*

$$
\|\mu^\pi_{M,h+1} - \mu^\pi_{\widetilde{M},h+1}\|_{\mathbb{TV}} \leq \mathbb{E}_{\pi,M}[\sum_{h'=1}^{h} L_\Gamma^{h-h'}\|\mathbb{P}_{T,h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'}) - \mathbb{P}_{\widetilde{T},h'}(\cdot|s_{h'},a_{h'},\mu^\pi_{M,h'})\|_{\mathbb{TV}}]. \quad (15)
$$

*Proof.* Under Assump. C, we can use a different way to decompose the density difference.

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

□

As implied by Lem. D.1 and Lem. D.2, we have the following corollary.

**Corollary D.3.** *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:*

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

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

*Moreover, with additional Assump. C, we have:*

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

**Theorem 6.2.** *[Model Difference Conversion] 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 $\pi$, under Assump. B, we have:*

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

*and*

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

*Proof.* By Assump. B, we have:

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

Then, by applying Corollary D.3, and plugging into the above equation, we can finish the proof. □

**Theorem 6.6.** *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 $\pi$, under Assump. B and Assump. C, we have:*

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

*Proof.* By applying Eq. (16) and Corollary D.3, 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 $\tau$ to be the smallest $t$ such that $\sum_{i=1}^{t}(X_i - 3Y_i) > C\log\frac{1}{\delta}$, we have:

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

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 $K \to +\infty$, 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 6.4.** *[Value Difference Lemma for MFC] Given an arbitrary model $M$ with transition function $\mathbb{P}_T$, and an arbitrary policy $\pi$, under Assump. B, we have:*

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

*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_{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_{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_{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_{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}}]. \qquad (17)
\end{aligned}
$$

we finish the proof by applying Corollary D.3. □

**Theorem D.5** (Result for MFC; Full Version of Thm. 5.1 and Thm. 5.2)**.** *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. 2, where $K$ is set to*

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

*with*

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

*or set to the following under additional Assump. C:*

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

*with*

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

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

*Proof.* On the event of Thm. 6.1, by Lem. 6.4, we have:

$$
\sum_{k=1}^{K} \mathcal{E}_{\mathrm{Opt}}(\pi^{k+1}) \leq \sum_{k=1}^{K} J_{M^{k+1}}(\pi^{k+1}) - J_{M^*}(\pi^{k+1}) \qquad (M^* \in \widehat{\mathcal{M}}^{k+1})
$$

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

Next, by applying Thm. 6.2 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. 2, for each $n\in[N]$, by Markov inequality, with probability at least $\frac{2}{3}$:

$$\begin{aligned}
&\mathcal{E}_{\text{Opt}}(\pi^{k_n}) = J_{M^*}(\pi^*_{\text{Opt}}) - J_{M^*}(\pi^{k_n}) && (18)\\
\le& 3\cdot\frac{1}{K}\cdot O\Big(H^2(1+L_TH)(1+L_rH)\frac{(1+L_T)^H-1}{L_T}\Big(\sqrt{K\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}} + \alpha HK\varepsilon_0\Big)\Big). && (19)\\
=& O\Big((1+L_TH)(1+L_rH)\frac{(1+L_T)^H-1}{L_T}\Big(\sqrt{\frac{1}{K}\text{dimE}_\alpha(\mathcal{M},\varepsilon_0)\log\frac{2|\mathcal{M}|KH}{\delta}} + \alpha H\varepsilon_0\Big)\Big). && (20)
\end{aligned}$$

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. (20) is less than $\frac{\varepsilon}{2}$. Therefore, in Alg. 2, with probability $1-\delta$, we have

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

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:

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

which implies

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

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

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

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

and we should choose

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

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

### D.4 Proofs for RL for Mean-Field Game

**Lemma 6.5.** *[Value Difference Lemma for MFG] Given two arbitrary model $M$ and $\widetilde{M}$, and two policies $\pi$ and $\widetilde{\pi}$, we have:*

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

$$
\begin{aligned}
&\qquad\qquad - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}] \\
&+ (2L_r H + 1)\mathbb{E}_{\pi,M}[\sum_{h=1}^H \|\mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu^\pi_{M,h}) \\
&\qquad\qquad - \mathbb{P}_{\widetilde{T},h}(\cdot|s_h, a_h, \mu^\pi_{\widetilde{M},h})\|_{\mathbb{TV}}]. \qquad (10)
\end{aligned}
$$

*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. (17) of Lem. 6.4, 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.3, and we finish the proof. □

**Theorem D.6** (Result for MFG; Full Version of Thm. 5.1 and Thm. 5.2)**.** *Under Assump. A and B, by running Alg. 1 with the MFG branch, after consuming $2HK$ trajectories, where $K$ is set to*

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

*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. C:*

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

*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 $\mathcal{E}_{\text{NE}}^{M}(\pi) := \max_{\widetilde{\pi}} \Delta_M(\widetilde{\pi}, \pi)$ to denote the exploitability in model $M$. Recall $M^{k+1}$ denotes the model such that $\pi^{k+1}$ is one of its equilibrium policies satisfying $\mathcal{E}_{\text{NE}}^{M^{k+1}}(\pi^{k+1}) = 0$. On the event in Thm. 6.1, $\forall k \in [K]$, we have $M^* \in \widehat{\mathcal{M}}^k$, which implies

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

By applying Lem. 6.5, Coro. D.3, and Thm. C.4, under Assump. B, we have:

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

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

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

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

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

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. C, we have:

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

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

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

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 $\mu_1$, and also for simplicity denote the deterministic expected rewards

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

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

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

where $\mu_1$ 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:

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

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 D** (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 Strict MFG-NE as a Fixed Point

We first introduce a stronger notion of NE, which we call the "Strict NE".

**Definition E.1** (Strict MFG-NE)**.** We call the policy $\pi^*$ a strict NE of if and only if the following holds for each $h, s$,

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

Note that a strict NE is always a NE. In the following, we only focus on the existence of strict NE.

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

$$Q_{\bar{h}}^\pi(s_{\bar{h}}, a_{\bar{h}}, \boldsymbol{\mu}) := \mathbb{E}\left[\sum_{h=\bar{h}}^H r_h(s_h, a_h, \mu_h)\middle| a_h \sim \pi_h(s_h), s_{h+1} \sim \mathbb{P}_{T,h}(\cdot|s_h, a_h, \mu_h),\ \forall\ h > \bar{h}\right]. \tag{21}$$

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 $\Gamma_{pop} : \{\Pi_h\}_{h\in[H]} \to \{\Delta(\mathcal{S}_h)\}_{h\in[H]}$ and $\Gamma_{pp} : \{\Pi_h\}_{h\in[H]} \times \{\Delta(\mathcal{S}_h)\}_{h\in[H]} \to \{\Pi_h\}_{h\in[H]}$ as

$$\Gamma_{pop}(\pi) := \{\mu_1\} \cup \{\mu_{h+1} := \underbrace{(\Gamma_{pop}(\pi_h, \ldots \Gamma_{pop,2}(\pi_2, \Gamma_{pop,1}(\pi_1, \mu_1)))}_{\text{from 1 to } h}\}_{h=1}^{H-1},$$
$$\Gamma_{pp}(\pi, \boldsymbol{\mu}) := \{\pi'_h(\cdot|s_h) := \arg\max_{u \in \Delta_{\mathcal{A}}} Q_h^\pi(s_h, \cdot, \boldsymbol{\mu})^\top u - \|\pi_h(\cdot|s_h) - u\|_2^2\}_{h=1}^H,$$

where $Q_h^\pi$ is the Q-value function defined in Eq. (21). The motivation for these operators is given by the following lemma:

**Lemma E.2** (Strict MFG-NE as fixed point)**.** *The tuple $\pi^*, \boldsymbol{\mu}^*$ is a strict MFG-NE if and only if the following conditions hold:*

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

*Proof.* First, assume $(\pi^*, \boldsymbol{\mu}^*)$ is a strict 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, $\Gamma_{SNE}(\pi^*) = \pi^*$.

Conversely, assume $\pi^* = \Gamma_{SNE}(\pi^*)$, that is, $\pi^*$ is a fixed point of the operator $\Gamma_{SNE}$. 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 $\Gamma_{pp}$ 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 $Q_h^{\pi^*}(s, \cdot, \boldsymbol{\mu}^*)^\top u - \|\pi_h(\cdot|s) - u\|_2^2$. We finish the proof. □

In the lemma above, the second condition is trivial to satisfy/compute once $\pi^*$ is known, hence the primary challenge will be in proving that the map $\Gamma_{SNE}$ 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. D is sufficient. The strategy will be to show that $\Gamma_{SNE}$ 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.3** (Continuity of $Q_h^\pi$)**.** *For any $s, a, h$, the map*

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

*is continuous.*

*Proof.* The proof follows from the fact that $Q_h^\pi$ 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.4** (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 (Nesterov et al., 2018). The Fenchel conjugate is useful due to the following well-known duality result.

**Lemma E.5.** *Assume that $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is differentiable and $\tau$-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 (Shalev-Shwartz and Singer, 2007) or Lemma 6.1.2 of (Nesterov et al., 2018). □

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

**Lemma E.6** (Proximal operator is non-expansive (Parikh et al., 2014))**.** *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.7** (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 $\Gamma_{pop}$ w.r.t. $\pi$ is straightforward by definition, as multiplications and additions of continuous functions are continuous.

For the continuity of $\Gamma_{pp}$, 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.6, 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.5, the map

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

is differentiable hence continuous for any $u \in \Delta_\mathcal{A}$, as the map $\|u - \cdot\|_2^2$ is weakly convex. By the continuity of $Q_h^\pi$ (see Lemma E.3), we can conclude that $\Gamma_{pp}$ 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.8** (Existence of MFG-NE; Formal Version of Prop. 3.2)**.** *Under Assump. D (which is implied by Assump. B), the map $\Gamma_{SNE}$ has a fixed point in the set $\{\Pi_h\}_{h\in[H]}$, that is, there exists a $\pi^*$ such that $\Gamma_{SNE}(\pi^*) = \pi^*$, and the tuple $(\pi^*, \Gamma_{pop}(\pi^*))$ is a strict MFG-NE, which implies the existence of NE.*

*Proof.* With the continuity of $\Gamma_{pop}, \Gamma_{pp}$, the know that the composition $\Gamma_{SNE}$ 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.2, this fixed point must constitute a strict MFG-NE.

Comparing with Eq. (4) and Def. E.1, we know the existence of strict NE implies the existence of NE. □