Learning Regularized Monotone Graphon Mean-Field Games

Fengzhuo Zhang[1]   **Vincent Y. F. Tan**[1]   **Zhaoran Wang**[2]   **Zhuoran Yang**[3]
[1]National University of Singapore   [2] Northwestern University   [3]Yale University
fzzhang@u.nus.edu, vtan@nus.edu.sg,
zhaoranwang@gmail.com, zhuoranyang.work@gmail.com

Abstract

This paper studies two fundamental problems in regularized Graphon Mean-Field Games (GMFGs). First, we establish the existence of a Nash Equilibrium (NE) of any λ-regularized GMFG (for $\lambda \geq 0$). This result relies on weaker conditions than those in previous works for analyzing both unregularized GMFGs ($\lambda = 0$) and λ-regularized MFGs, which are special cases of GMFGs. Second, we propose provably efficient algorithms to learn the NE in weakly monotone GMFGs, motivated by Lasry and Lions [2007]. Previous literature either only analyzed continuous-time algorithms or required extra conditions to analyze discrete-time algorithms. In contrast, we design a discrete-time algorithm and derive its convergence rate solely under weakly monotone conditions. Furthermore, we develop and analyze the action-value function estimation procedure during the online learning process, which is absent from algorithms for monotone GMFGs. This serves as a sub-module in our optimization algorithm. The efficiency of the designed algorithm is corroborated by empirical evaluations.

1 Introduction

In Multi-Agent Reinforcement Learning (MARL), the sizes of state and action spaces grow exponentially in the number of agents, which is known as the "curse of many agents" [Sonu et al., 2017, Wang et al., 2020] and restrict its applicability to large-scale scenarios. The Mean-Field Game (MFG) has thus been proposed to mitigate this problem by exploiting the homogeneity assumption [Huang et al., 2006, Lasry and Lions, 2007], and it has achieved tremendous successes in many real-world applications [Cousin et al., 2011, Achdou et al., 2020]. However, the homogeneity assumption is an impediment when modeling scenarios in which the agents are heterogeneous. GMFGs, as extensions of MFGs, have thus been proposed to model the behaviors of heterogeneous agents and ameliorate the "curse of many agents" at the same time [Parise and Ozdaglar, 2019, Carmona et al., 2022].

Despite the empirical successes of the Graphon Mean-Field Game (GMFG) [Aurell et al., 2022a], its theoretical understanding is lacking. First, sufficient conditions for Nash Equilibrium (NE) existence in regularized GMFG have not been established. Most works only address the existence of the NE in *unregularized* GMFGs. However, *regularization* is employed in practical implementations for improved exploration and robustness [Geist et al., 2019]. Moreover, previous works prove the existence of NE in regularized MFGs, a special case of GMFGs, only under the *contraction condition*, which is overly restrictive for real-world applications. Second, the analysis of *discrete-time* algorithms for monotone GMFGs is lacking. Most existing works design provably efficient discrete-time algorithms only under contraction conditions, as shown in Table 1. Complementarily, previous works on monotone GMFGs either only derive the convergence rate for continuous-time algorithms, which ignores the discretization error, or require extra conditions, (e.g., potential games) to analyze discrete-time algorithms.

Table 1: Comparison of GMFGs and MFGs learning algorithms

	Condition	No population manipulation	Online playing	Heterogeneity	Discrete-time algorithm	Convergence rate
Anahtarci et al. [2022]	Contraction	No	No	No	Yes	Yes
Xie et al. [2021a]	Contraction	No	Yes	No	Yes	Yes
Zaman et al. [2022]	Contraction	No	Yes	No	Yes	Yes
Yardim et al. [2022]	Contraction	Yes	Yes	No	Yes	Yes
Perrin et al. [2020]	Monotone	No	No	No	No	Yes
Geist et al. [2021]	Potential &Monotone	No	No	No	Yes	Yes
Perolat et al. [2021]	Monotone	Yes	Yes	Yes	No	No
Fabian et al. [2022]	Monotone	Yes	Yes	Yes	No	No
Our work	**Monotone**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

In this paper, we first consider GMFGs in full generality, i.e., without any contraction or monotone conditions. The goal is to establish the existence of NE in the regularized GMFG in this general setting. Then we focus on monotone GMFGs motivated by Lasry and Lions [2007]. We aim to learn the unique NE from the online interactions of all agents with and without the action-value function oracle. When the oracle is absent, the action-value functions should be estimated from the data of sampled agents generated in the online game.

In the analysis, we have to overcome difficulties that arise from both the *existence* of the NE problem and the *learning* of the NE. First, the proof of the existence of NE in regularized GMFG involves establishing some topological spaces and operators related to NE on which fixed point theorems are applicable. However, the direct construction of the space and the operators for GMFGs with uncountably infinite agents is challenging. Second, the design and analysis of the discrete-time NE learning algorithm require subtle exploitation of the monotone condition. Unlike continuous-time algorithms with infinitesimal step sizes, the design of appropriate step sizes is additionally required for the discrete-time algorithm to guarantee that iterates evolve appropriately. This guarantee originates from the delicate utilization of the monotone condition in the optimization procedures.

To address the difficulty of the existence problem, we construct a regularized MFG from the regularized GMFG and show that the NE of the constructed game can be converted into the NE of the original game, thus mitigating the difficulty of having an uncountable number of agents. To handle the difficulty in the NE learning problem, we design the Monotone GMFG Policy Mirror Descent (MonoGMFG-PMD) algorithm, which iteratively implements policy mirror descent for each player. We show that this procedure results in a decrease of the KL divergence between the iterate and the NE, and this decrease is related to the gap appearing in the weakly monotone condition. When the action-value function oracle is absent, we also design and analyze action-value functions estimation algorithms to serve as a submodule of the optimization procedures.

Main Contributions: We first establish the existence of the NE in the λ-regularized GMFG with $\lambda \geq 0$ assuming Lipschitzness of graphons and continuity of transition kernels and reward functions. Our result relaxes the assumption of the Lipschitzness of transition kernels and rewards required in previous works on unregularized GMFGs and the contraction condition in the literature on regularized MFG [Cui and Koeppl, 2021a]. Then we design and analyze the MonoGMFG-PMD algorithm. Using an action-value function oracle, the convergence rate for MonoGMFG-PMD is proved to be $\tilde{O}(T^{-1/2})$ after T iterations. Without the oracle, the convergence rate includes an additional $\tilde{O}(K^{-1/2} + N^{-1})$ term that arises from sampling N agents and collecting data from K episodes, reflecting the generalization error and the approximation error of the estimation algorithm. As shown in Table 1 , our algorithm can be implemented from the online interaction of agents and does not require the distribution flow manipulation. Detailed explanations of the properties stated in the columns of Table 1 are provided in Appendix A. Our result for MonoGMFG-PMD provides the *first convergence rate for discrete-time algorithms in monotone GMFGs*.

2   Related Works

MFGs were proposed by Huang et al. [2006] and Lasry and Lions [2007] to model the interactions among a set of homogeneous agents. In recent years, learning the NE of the MFGs formulated by

discrete-time Markov Decision Process (MDP)s has attracted a lot of interest. There is a large body of works that design and analyze algorithms for the MFGs under contraction conditions [Anahtarci et al., 2019, 2022, Cui and Koeppl, 2021a, Xie et al., 2021a, Zaman et al., 2022, Yardim et al., 2022]. Typically, these works design reinforcement learning algorithms to approximate the contraction operators in MFGs, and the NE is learned by iteratively applying this operator. In contrast, another line of works focuses on the MFGs under monotone conditions. Motivated by Lasry and Lions [2007], the transition kernels in these works are independent of the mean fields. Perrin et al. [2020] proposed and analyzed the continuous-time fictitious play algorithm, which dynamically weighs the past mean fields and the best response policies to derive new mean fields and policies. With the additional potential assumption, Geist et al. [2021] derived the convergence rate for the discrete-time fictitious play algorithm. Perolat et al. [2021] then proposed the continuous-time policy mirror descent algorithm but only proved the asymptotic convergence, i.e., the consistency. In summary, there is no convergence rate result for any discrete-time algorithm for MFGs under the monotone condition. In addition, the relationship between the contraction conditions and the monotone conditions is not clear from existing works, but they complement each other.

To capture the potential heterogeneity among agents, GMFGs have been proposed by Parise and Ozdaglar [2019] in static settings as an extension of MFGs. The heterogeneous interactions among agents are represented by graphons. Aurell et al. [2022b], Caines and Huang [2019] then extended the GMFGs to the continuous-time setting, where the existence and the uniqueness of NE were established. Vasal et al. [2020] formulated discrete-time GMFGs and provided way to compute the NE with master equations. With the precise underlying graphons values, Cui and Koeppl [2021b] proposed algorithms to learn the NE of GMFGs with the contraction condition by modifying MFGs learning algorithms. Fabian et al. [2022] considered the monotone GMFG and proposed the continuous-time policy mirror descent algorithm to learn the NE. However, only consistency was provided in the latter two works.

Notation Let $[N] := \{1, \cdots, N\}$. Given a measurable space $(\Omega, \mathcal{F})$, we denote the collection of all the measures and the probability measures on $(\Omega, \mathcal{F})$ as $\mathcal{M}(\Omega)$ and $\Delta(\Omega)$, respectively. For a metric space $(\mathcal{X}, \|\cdot\|)$, we use $C(\mathcal{X})$ and $C_b(\mathcal{X})$ to denote the set of all continuous functions and the set of all bounded continuous functions on $\mathcal{X}$, respectively. For a measurable space $(\mathcal{X}, \mathcal{F})$ and two distributions $P, Q \in \Delta(\mathcal{X})$, the total variation distance between them is defined as $\mathrm{TV}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$. A sequence of measures $\{\mu_n\}$ on $\mathcal{X}$ is said to *converge weakly* to a measure μ if $\int_{\mathcal{X}} g(x)\mu_n(\mathrm{d}x) \to \int_{\mathcal{X}} g(x)\mu(\mathrm{d}x)$ for all $g \in C_b(\mathcal{X})$.

3 Preliminaries

3.1 Graphon Mean-Field Games

We consider a GMFG that is defined through a tuple $(\mathcal{I}, \mathcal{S}, \mathcal{A}, H, P, r, W, \mu_1)$. The horizon (or length) of the game is denoted as H. In GMFG, we have infinite players, each corresponding to a point $\alpha \in \mathcal{I} = [0, 1]$. The state and action space of them are the same, denoted as $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ respectively. The interaction among players is captured by *graphons*. Graphons are symmetric functions that map $[0, 1]^2$ to $[0, 1]$. Symmetry here refers to that $W(\alpha, \beta) = W(\beta, \alpha)$ for all $\alpha, \beta \in [0, 1]$. We denote the set of graphons as $\mathcal{W} = \{W : [0, 1]^2 \to [0, 1] \mid W \text{ is symmetric.}\}$. The set of graphons of the game is $W = \{W_h\}_{h=1}^H$ with $W_h \in \mathcal{W}$. The state transition and reward of each player are influenced by the collective behavior of all the other players through an *aggregate* $z \in \mathcal{M}(\mathcal{S})$. At time $h \in [H]$, we denote the state distribution of player $\beta \in \mathcal{I}$ as $\mu_h^\beta \in \Delta(\mathcal{S})$. The aggregate for player $\alpha \in \mathcal{I}$ with the underlying graphon $W_h \in \mathcal{W}$ is then defined as

$$z_h^\alpha = \int_0^1 W_h(\alpha, \beta)\mu_h^\beta \,\mathrm{d}\beta. \tag{1}$$

The transition kernels $P = \{P_h\}_{h=1}^H$ of the game are functions $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{M}(\mathcal{S}) \to \mathcal{S}$ for all $h \in [H]$. At time h, if player α takes action $a_h^\alpha \in \mathcal{A}$ at state $s_h^\alpha \in \mathcal{S}$, her state will transition according to $s_{h+1}^\alpha \sim P_h(\cdot \mid s_h^\alpha, a_h^\alpha, z_h^\alpha)$. The reward functions are denoted as $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{M}(\mathcal{S}) \to [0, 1]$ for all $h \in [H]$. We note that the players in GMFG are heterogeneous. This means that different players will, in general, receive different aggregates from other players. The distribution $\mu_1 \in \Delta(\mathcal{S})$ is the initial state distribution for all the players. A policy for an player α is $\pi^\alpha = \{\pi_h^\alpha\}_{h=1}^H \in \Pi^H$, where $\pi_h^\alpha : \mathcal{S} \to \Delta(\mathcal{A})$ takes action based only on the current state, and Π is the set of all these policies.

A policy for all the players $\pi^{\mathcal{I}} \in \tilde{\Pi} = \Pi^{H \times \mathcal{I}}$ is the collection of the policies of each player, i.e, $\pi^{\mathcal{I}} = \{\pi^{\alpha}\}_{\alpha \in \mathcal{I}}$. In the following, we denote the state distributions of all the players at time h and the state distributions of all the players at any time (distribution flow) respectively as $\mu_h^{\mathcal{I}} = \{\mu_h^{\alpha}\}_{\alpha \in \mathcal{I}}$ and $\mu^{\mathcal{I}} = \{\mu_h^{\mathcal{I}}\}_{h=1}^{H} \in \tilde{\Delta} = \Delta(\mathcal{S})^{H \times \mathcal{I}}$. Eqn. (1) shows that the aggregate z_h^{α} is a function of $\mu_h^{\mathcal{I}}$ and W_h, so to make this dependence explicit, we also write it as $z_h^{\alpha}(\mu_h^{\mathcal{I}}, W_h)$.

We consider the entropy-regularized GMFG. It has been shown that the regularization results in policy gradient algorithms converging faster [Shani et al., 2020, Cen et al., 2022]. In this game, the rewards of each player are the sum of the original rewards and the negative entropy of the policy multiplied by a constant. In a λ-regularized GMFG ($\lambda \geq 0$), the value function and the action-value function of player α with policy π^{α} on the MDP induced by the distribution flow $\mu^{\mathcal{I}}$ are defined as

$$V_h^{\lambda,\alpha}(s, \pi^{\alpha}, \mu^{\mathcal{I}}) = \mathbb{E}^{\pi^{\alpha}}\left[\sum_{t=h}^{H} r_t(s_t^{\alpha}, a_t^{\alpha}, z_t^{\alpha}(\mu_t^{\mathcal{I}}, W_t)) - \lambda \log \pi_t^{\alpha}(a_t^{\alpha} \mid s_t^{\alpha}) \,\Big|\, s_h^{\alpha} = s\right], \qquad (2)$$

$$Q_h^{\lambda,\alpha}(s, a, \pi^{\alpha}, \mu^{\mathcal{I}}) = r_h(s, a, z_h^{\alpha}(\mu_h^{\mathcal{I}}, W_h)) + \mathbb{E}^{\pi^{\alpha}}\left[V_{h+1}^{\lambda,\alpha}(s_{h+1}^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) \mid s_h^{\alpha} = s, a_h^{\alpha} = a\right] \qquad (3)$$

for all $h \in [H]$, where the expectation $\mathbb{E}^{\pi^{\alpha}}$ is taken with respect to the stochastic process induced by implementing policy π^{α} on the MDP induced by $\mu^{\mathcal{I}}$. The cumulative reward function of player α is defined as $J^{\lambda,\alpha}(\pi^{\alpha}, \mu^{\mathcal{I}}) = \mathbb{E}_{\mu_1}[V_1^{\lambda,\alpha}(s, \pi^{\alpha}, \mu^{\mathcal{I}})]$. Then the notion of an NE is defined as follows.

Definition 1. *An NE of the λ-regularized GMFG is a pair $(\pi^{*,\mathcal{I}}, \mu^{*,\mathcal{I}}) \in \tilde{\Pi} \times \tilde{\Delta}$ that satisfies: (i) (player rationality) $J^{\lambda,\alpha}(\pi^{*,\alpha}, \mu^{*,\mathcal{I}}) = \max_{\tilde{\pi}^{\alpha} \in \Pi^H} J^{\lambda,\alpha}(\tilde{\pi}^{\alpha}, \mu^{*,\mathcal{I}})$ for all $\alpha \in \mathcal{I}$ up to a zero-measure set on $\mathcal{I}$ with respect to the Lebesgue measure. (ii) (Distribution consistency) The distribution flow $\mu^{*,\mathcal{I}}$ is equal to the distribution flow induced by implementing the policy $\pi^{*,\mathcal{I}}$.*

Similar to the NE for the finite-player games, the NE of the λ-regularized GMFG requires that the policy of each player is optimal. However, in GMFGs, the optimality is with respect to $\mu^{*,\mathcal{I}}$.

3.2 Mean-Field Games

As an important subclass of GMFG, MFG corresponds to the GMFG with *constant* graphons, i.e, $W(\alpha, \beta) = p$ for all $\alpha, \beta \in \mathcal{I}$. MFGs involve infinite *homogeneous* players. All the players employ the same policy and thus share the same distribution flow. The aggregate in Eqn. 1 degenerates to $z_h^{\alpha} = \int_0^1 p \cdot \mu_h^{\beta} \mathrm{d}\beta = p \cdot \mu_h$ for all $\alpha \in \mathcal{I}$. Here μ_h is the state distribution of a representative player. Thus, an MFG is denoted by a tuple $(\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{H}, \bar{P}, \bar{r}, \mu_1)$. The state space, the action space, and the horizon are respectively denoted as $\bar{\mathcal{S}}$, $\bar{\mathcal{A}}$, and $\bar{H}$. Here, the transition kernels $\bar{P} = \{\bar{P}_h\}_{h=1}^{H}$ are functions $\bar{P}_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \to \mathcal{S}$, and reward functions $\bar{r}_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \to [0, 1]$ for all $h \in [H]$. In the MFG, all the players adopt the *same* policy $\pi = \{\pi_h\}_{h=1}^{H}$ where $\pi_h \in \Pi$. The value function and the action-value function in the λ-regularized MFG with the underlying distribution flow $\mu = \{\mu_h\}_{h=1}^{H} \in \Delta(\mathcal{S})^H$ can be similarly defined as Eqn. (2) and (3) respectively as follows

$$\bar{V}_h^{\lambda}(s, \pi, \mu) = \mathbb{E}^{\pi}\left[\sum_{t=h}^{H} \bar{r}_t(s_t, a_t, \mu_t) - \lambda \log \pi_t(a_t \mid s_t) \,\Big|\, s_h = s\right],$$

$$\bar{Q}_h^{\lambda}(s, a, \pi, \mu) = \bar{r}_h(s, a, \mu_h) + \mathbb{E}^{\pi}\left[\bar{V}_{h+1}^{\lambda}(s_{h+1}, \pi, \mu) \mid s_h = s, a_h = a\right]$$

for all $h \in [H]$. The cumulative reward is $\bar{J}^{\lambda}(\pi, \mu) = \mathbb{E}_{\mu_1}[\bar{V}_1^{\lambda}(s, \pi, \mu)]$. The notion of NE can be similarly derived as follows.

Definition 2. *An NE of the λ-regularized MFG is a pair $(\pi^*, \mu^*) \in \Pi^H \times \Delta(\mathcal{S})^H$ that satisfies: (i) (player rationality) $J^{\lambda}(\pi^*, \mu^*) = \max_{\tilde{\pi} \in \Pi^H} J^{\lambda}(\tilde{\pi}, \mu^*)$. (ii) (Distribution consistency) The distribution flow μ^* is equal to the distribution flow induced by the policy π^*.*

Remark 1. *Compared with Definition 1, the definition of NE in MFG only involves the policy and the distribution flow of a single representative player, since the agents are homogeneous in MFGs.*

4 Existence of the NEs in Regularized GMFGs and MFGs

We now state some assumptions to demonstrate the existence of a NE for λ-regularized GMFGs.

Assumption 1. *The state space $\mathcal{S}$ is compact, and the action space $\mathcal{A}$ is finite.*

This assumption imposes rather mild constraints on $\mathcal{S}$ and $\mathcal{A}$. In real-world applications, the states are usually physical quantities and thus reside in a compact set. For the action space, many deep reinforcement learning algorithms discretize the potential continuous action sets into finite sets [Lowe et al., 2017, Mordatch and Abbeel, 2018].

Assumption 2. *The graphons W_h for $h \in [H]$ are continuous functions.*

The stronger version of this assumption (Lipschitz continuity) is widely adopted in GMFG works [Cui and Koeppl, 2021b, Fabian et al., 2022]. It helps us to build the continuity of the transition kernels and rewards with respect to players.

Assumption 3. *For all $h \in [H]$, the reward function $r_h(s, a, z)$ is continuous on $\mathcal{S} \times \mathcal{A} \times \mathcal{M}(\mathcal{S})$, that is if $(s_n, a_n, z_n) \to (s, a, z)$ as $n \to \infty$, then $r_h(s_n, a_n, z_n) \to r_h(s, a, z)$. The transition kernel $P_h(\cdot \,|\, s, a, z)$ is weakly continuous in $\mathcal{S} \times \mathcal{A} \times \mathcal{M}(\mathcal{S})$, that is if $(s_n, a_n, z_n) \to (s, a, z)$ as $n \to \infty$, $P_h(\cdot \,|\, s_n, a_n, z_n) \to P_h(\cdot \,|\, s, a, z)$ weakly.*

This assumption states the continuity of the models, i.e., the transition kernels and the rewards, as functions of the state, action, and aggregate. We note that the Lipschitz continuity assumption of the model in the previous works implies that our assumption is satisfied [Cui and Koeppl, 2021b, Fabian et al., 2022]. Next, we state the existence result of regularized GMFG.

Theorem 1. *Under Assumptions 1, 2 and 3, for all $\lambda \geq 0$, the λ-regularized GMFG $(\mathcal{I}, \mathcal{S}, \mathcal{A}, H, P, r, W, \mu_1)$ admits an NE $(\pi^{\mathcal{I}}, \mu^{\mathcal{I}}) \in \tilde{\Pi} \times \tilde{\Delta}$.*

This theorem strengthens previous existence results in Cui and Koeppl [2021b] and Fabian et al. [2022] in two aspects. First, our assumptions are weaker. These two existing works require a finite state space and the Lipschitz continuity of the models. In contrast, we only need a compact state space and the model continuity. Second, their results only hold for the unregularized case ($\lambda = 0$), whereas ours holds for any $\lambda \geq 0$. In the proof of Theorem 1, we construct a MFG from the given GMFG and show that an NE of the constructed MFG can be converted to an NE of the GMFG. Then we prove the existence of NE in the constructed regularized MFG. Our existence result for the regularized MFG is also a significant addition to the MFG literature.

Remark 2. *Although we show that an NE of the constructed MFG can be converted to an NE of GMFG, this proof does not imply that GMFG forms a subclass of or is equivalent to MFG. This is because we have only demonstrated the relationship between the NEs of these two games, but the exact realizations of the GMFG and the conceptually constructed MFG may differ. It means that the sample paths of these two games may not be the same, which include the realizations of the states, actions, and rewards of all the players.*

We next state the assumption needed for the MFG.

Assumption 4. *The MFG $(\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{H}, \bar{P}, \bar{r}, \mu_1)$ satisfies that: (i) The state space $\bar{\mathcal{S}}$ is compact, and the action space $\bar{\mathcal{A}}$ is finite. (ii) The reward functions $\bar{r}_h(s, a, \mu)$ for $h \in [H]$ are continuous on $\mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$. The transition kernels are weakly continuous on $\mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$; that is if $(s_n, a_n, \mu_n) \to (s, a, \mu)$ as $n \to \infty$, $\bar{P}_h(\cdot \,|\, s_n, a_n, \mu_n) \to \bar{P}_h(\cdot \,|\, s, a, \mu)$ weakly.*

Then the existence of the NE is stated as follows.

Theorem 2. *Under Assumption 4, the λ-regularzied MFG $(\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{H}, \bar{P}, \bar{r}, \mu_1)$ admits an NE $(\pi, \mu) \in \Pi^H \times \Delta(\mathcal{S})^H$ for any $\lambda \geq 0$.*

Our result in Theorem 2 imposes weaker conditions than previous works [Cui and Koeppl, 2021a, Anahtarci et al., 2022] to guarantee the existence of an NE. These existing works prove the existence of NE by assuming a *contractive* property and the finiteness of the state space. They also require a strong Lispchitz assumption [Anahtarci et al., 2022], where the Lipschitz constants of the models should be small enough. In contrast, we only require the continuity assumption in Theorem 2. This is due to our analysis of the operator for the regularized MFG and the use of Kakutani fixed point theorem Guide [2006].

5 Learning NE of Monotone GMFGs

In this section, we focus on GMFGs with transition kernels that are independent of the aggregate z, i.e., $P_h : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ for $h \in [H]$. This model is motivated by the seminal work Lasry and Lions [2007], where the state evolution in continuous-time MFG is characterized by the Fokker–Plank equation. However, the form of the Fokker–Plank equation results in the state transition of each player being independent of other players. This model is also widely accepted in the discrete-time GMFG and MFG literature [Fabian et al., 2022, Perrin et al., 2020, Perolat et al., 2021].

5.1 Monotone GMFG

We first generalize the notion of *monotonicity* from multi-population MFGs in Perolat et al. [2021] to GMFGs.

Definition 3 (Weakly Monotone Condition). *A GMFG is said to be* weakly monotone *if for any $\rho^{\mathcal{I}}, \tilde{\rho}^{\mathcal{I}} \in \Delta(\mathcal{S} \times \mathcal{A})^{\mathcal{I}}$ and their marginalizations on the states $\mu^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}} \in \Delta(\mathcal{S})^{\mathcal{I}}$, we have*

$$\int_{\mathcal{I}} \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \left(\rho^{\alpha}(s, a) - \tilde{\rho}^{\alpha}(s, a) \right) \left(r_h \left(s, a, z_h^{\alpha}(\mu^{\mathcal{I}}, W_h) \right) - r_h \left(s, a, z_h^{\alpha}(\tilde{\mu}^{\mathcal{I}}, W_h) \right) \right) \mathrm{d}s \, \mathrm{d}\alpha \leq 0$$

for all $h \in [H]$, where W_h is the underlying graphon. It is strictly weakly monotone *if the inequality is strict when $\rho^{\mathcal{I}} \neq \tilde{\rho}^{\mathcal{I}}$.*

In two MDPs induced by the distribution flows $\mu^{\mathcal{I}}$ of $\pi^{\mathcal{I}}$ and $\tilde{\mu}^{\mathcal{I}}$ of $\tilde{\pi}^{\mathcal{I}}$, the weakly monotone condition states that we can achieve higher rewards at stage $h \in [H]$ by swapping the policies. This condition has two important implications. The first is the uniqueness of the NE.

Proposition 1. *Under Assumptions 1, 2, and 3, a strictly weakly monotone λ-regularized GMFG has a unique NE for any $\lambda \geq 0$ up to a zero-measure set on $\mathcal{I}$ with respect to the Lebesgue measure.*

In the following, we denote this unique NE as $(\pi^{*, \mathcal{I}}, \mu^{*, \mathcal{I}})$, and we aim to learn this NE. The second implication concerns the relationship between the cumulative rewards of two policies.

Proposition 2. *If a λ-regularized GMFG satisfies the weakly monotone condition, then for any two policies $\pi^{\mathcal{I}}, \tilde{\pi}^{\mathcal{I}} \in \tilde{\Pi}$ and their induced distribution flows $\mu^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}} \in \tilde{\Delta}$, we have*

$$\int_0^1 J^{\lambda, \alpha}(\pi^{\alpha}, \mu^{\mathcal{I}}) + J^{\lambda, \alpha}(\tilde{\pi}^{\alpha}, \tilde{\mu}^{\mathcal{I}}) - J^{\lambda, \alpha}(\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}) - J^{\lambda, \alpha}(\pi^{\alpha}, \tilde{\mu}^{\mathcal{I}}) \, \mathrm{d}\alpha \leq 0$$

If the λ-regularized GMFG satisfies the strictly weakly monotone condition, then the inequality is strict when $\pi^{\mathcal{I}} \neq \tilde{\pi}^{\mathcal{I}}$.

Proposition 2 shows that if we have two policies, we can improve the cumulative rewards on the MDP induced by these policies by swapping the policies. This implies an important property of the NE $(\pi^{*, \mathcal{I}}, \mu^{*, \mathcal{I}})$. Since $\pi^{*, \mathcal{I}}$ is optimal on the MDP induced by $\mu^{*, \mathcal{I}}$, we have $\int_0^1 J^{\lambda, \alpha}(\pi^{*, \alpha}, \mu^{*, \mathcal{I}}) \, \mathrm{d}\alpha \geq \int_0^1 J^{\lambda, \alpha}(\pi^{\alpha}, \mu^{*, \mathcal{I}}) \, \mathrm{d}\alpha$ for any $\pi^{\mathcal{I}} \in \tilde{\Pi}$. Then Proposition 2 shows that $\int_0^1 J^{\lambda, \alpha}(\pi^{*, \alpha}, \mu^{\mathcal{I}}) \, \mathrm{d}\alpha \geq \int_0^1 J^{\lambda, \alpha}(\pi^{\alpha}, \mu^{\mathcal{I}}) \, \mathrm{d}\alpha$ for any policy $\pi^{\mathcal{I}}$ and the distribution flow $\mu^{\mathcal{I}}$ it induces. This means that the NE policy gains cumulative rewards not less than any policy $\pi^{\mathcal{I}}$ on the MDP induced by $\pi^{\mathcal{I}}$. This motivates the design of our NE learning algorithm.

5.2 Policy Mirror Descent Algorithm for Monotone GMFG

In this section, we introduce the algorithm to learn the NE, which is called MonoGMFG-PMD and whose pseudo-code is outlined in Algorithm 1. It consists of three main steps. The first step estimates the action-value function (Line 3). We need to evaluate the action-value function of a policy on the MDP induced by itself. This estimate can be obtained for each player independently by playing the $\pi_t^{\mathcal{I}}$ several times. We assume access to a sub-module for this and quantify the estimation error in our analysis. The second step is the policy mirror descent (Line 4). Given $\lambda \eta_t < 1$, This step can be equivalently formulated as

$$\hat{\pi}_{t+1, h}^{\alpha}(\cdot \,|\, s) = \operatorname*{argmax}_{p \in \Delta(\mathcal{A})} \frac{\eta_t}{1 - \lambda \eta_t} \left[\langle \hat{Q}_h^{\lambda, \alpha}(s, \cdot, \pi_t^{\alpha}, \mu_t^{\mathcal{I}}), p \rangle - \lambda R(p) \right] - D_{\mathrm{kl}} \left(p \| \pi_{t, h}^{\alpha}(\cdot \,|\, s) \right) \quad \forall s \in \mathcal{S},$$

Algorithm 1 MonoGMFG-PMD

Procedure:
 1: Initialize $\pi_{1,h}^{\alpha}(\cdot \mid s) = \mathrm{Unif}(\mathcal{A})$ for all $s \in \mathcal{S}, h \in [H]$ and $\alpha \in \mathcal{I}$.
 2: **for** $t = 1, 2, \cdots, T$ **do**
 3:     Compute the action-value function $\hat{Q}_h^{\lambda,\alpha}(s, a, \pi_t^{\alpha}, \mu_t^{\mathcal{I}})$ for all $\alpha \in \mathcal{I}$ and $h \in [H]$, where $\mu_t^{\mathcal{I}}$ is
     the distribution flow induced by $\pi_t^{\mathcal{I}}$.
 4:     $\hat{\pi}_{t+1,h}^{\alpha}(\cdot \mid s) \propto \left(\pi_{t,h}^{\alpha}(\cdot \mid s)\right)^{1-\lambda\eta_t} \exp\left(\eta_t \hat{Q}_h^{\lambda,\alpha}(s, a, \pi_t^{\alpha}, \mu_t^{\mathcal{I}})\right)$ for all $\alpha \in \mathcal{I}$ and $h \in [H]$
 5:     $\pi_{t+1,h}^{\alpha}(\cdot \mid s) = (1 - \beta_t)\hat{\pi}_{t+1,h}^{\alpha}(\cdot \mid s) + \beta_t\mathrm{Unif}(\mathcal{A})$
 6: **end for**
 7: Output $\bar{\pi}^{\mathcal{I}} = \mathrm{Unif}\left(\pi_{[1:T]}^{\mathcal{I}}\right)$

---

where $R(p) = \langle p, \log p \rangle$ is the negative entropy function. This step aims to improve the performance of the policy $\pi_t^{\mathcal{I}}$ on its own induced MDP. Intuitively, since the policy $\pi^{*,\mathcal{I}}$ in NE performs better than $\pi_t^{\mathcal{I}}$ on the MDP induced by $\mu_t^{\mathcal{I}}$ as shown in Section 5.1, the improved policy $\pi_{t+1}^{\mathcal{I}}$ should be closer to $\pi^{*,\mathcal{I}}$ than $\pi_t^{\mathcal{I}}$. The third step mixes the current policy with the uniform policy (Line 5) to encourage exploration.

MonoGMFG-PMD is different from previous NE learning algorithms for monotone GMFG in Perolat et al. [2021], Fabian et al. [2022] in three different ways. First, MonoGMFG-PMD is designed to learn the NE of the λ-regularized GMFG with $\lambda > 0$, whereas other algorithms learn the NE of the *unregularized* GMFGs. As a result, our policy improvement (Line 4) discounts the previous policy as $(\pi_{t,h}^{\alpha})^{1-\lambda\eta_t}$, but other algorithms retain $\pi_{t,h}^{\alpha}$. Second, our algorithm is discrete-time and thus amenable for practical implementation. However, other provably efficient algorithms evolve in continuous time. Finally, MonoGMFG-PMD encourages exploration in Line 5, which is important for the theoretical analysis. Such a step is missing in other algorithms.

5.3 Theoretical Analysis for MonoGMFG-PMD with Estimation Oracle

This section provides theoretical analysis for the MonoGMFG-PMD algorithm given an action-value function oracle in Line 3, i.e., $\hat{Q}_h^{\lambda,\alpha} = Q_h^{\lambda,\alpha}$. We denote the unique NE of the λ-regularized GMFG as $(\pi^{*,\mathcal{I}}, \mu^{*,\mathcal{I}})$. For any policy $\pi^{\mathcal{I}}$, we measure the distance to the policy $\pi^{*,\mathcal{I}}$ of NE using

$$D(\pi^{\mathcal{I}}) = \int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}\left[D_{\mathrm{kl}}\left(\pi_h^{*,\alpha}(\cdot \mid s_h^{\alpha}) \| \pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\right)\right] \mathrm{d}\alpha.$$

This metric measures the weighted KL divergence between policy $\pi^{\mathcal{I}}$ and the NE policy, and the weights are the NE distribution flow $\mu^{*,\mathcal{I}}$.

Theorem 3. *Assume that the GMFG is strictly weakly monotone and we have an action-value function oracle. Let $\eta_t = \eta = O(T^{-1/2})$ and $\beta_t = \beta = O(T^{-1})$ in MonoGMFG-PMD. For any $\lambda > 0$ we have*

$$D\left(\frac{1}{T}\sum_{t=1}^T \pi_t^{\mathcal{I}}\right) = O\left(\frac{\lambda \log^2 T}{T^{1/2}}\right).$$

Theorem 3 provides the first convergence rate result for a discrete-time algorithm on strictly weakly monotone GMFGs under mild assumptions. In contrast, Perolat et al. [2021], Fabian et al. [2022] only consider the continuous-time policy mirror descent, which is difficult for the practical implementation, and only provide the asymptotic consistency results. Geist et al. [2021] derive exploitability results for a fictitious play algorithm but require the potential structure and the Lipschitzness of the reward function. Our proof for Theorem 3 mainly exploits properties of NE discussed in Section 5.1. Concretely, we use the fact that the policy mirror descent procedure reduces the distance between the policy iterate and the NE as $D(\pi_{t+1}^{\mathcal{I}}) - D(\pi_t^{\mathcal{I}}) \approx \int_0^1 J^{\lambda,\alpha}(\pi_t^{\alpha}, \mu_t^{\mathcal{I}}) - J^{\lambda,\alpha}(\pi^{*,\alpha}, \mu_t^{\mathcal{I}}) \mathrm{d}\alpha$. Thus, the policy iterate becomes closer to the NE. However, the discretization error and the exploration influence (Line 5) also appear, requiring additional care to show that $D(\pi_t^{\mathcal{I}})$, in general, decreases.

Algorithm 2 Estimation of Action-value Function

Procedure:

1: Sample N players $\{i/N\}_{i=1}^{N} \subseteq [0,1]$.
2: The ith player implements $\pi_t^{\text{b},i}$ for $i \in [N]$, and the other players implement $\pi_t^{\mathcal{I}}$.
3: Collect data $\{(s_{\tau,h}^i, a_{\tau,h}^i, r_{\tau,h}^i)\}_{i,\tau,h=1}^{N,K,H}$ of sampled players from K independent episodes.
4: Initialize $\hat{V}_{H+1}^{\lambda,i}(s,a) = 0$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $i \in [N]$.
5: **for** time $h = H, \cdots, 1$ **do**
6:     **for** Player $i = 1, \cdots, N$(in parallel) **do**
7:         $\hat{Q}_h^{\lambda,i} = \operatorname{argmin}_{f \in \mathcal{F}_h} \sum_{\tau=1}^{K} \left(f(s_{\tau,h}^i, a_{\tau,h}^i) - r_{\tau,h}^i - \hat{V}_{h+1}^{\lambda,i}(s_{\tau,h+1}^i) \right)^2$.
8:         $\hat{V}_h^{\lambda,i}(s) = \langle \hat{Q}_h^{\lambda,i}(s,\cdot), \pi_{t,h}^{i/N}(\cdot,s) \rangle - \lambda R\big(\pi_{t,h}^{i/N}(\cdot,s)\big)$.
9:     **end for**
10: **end for**
11: Output $\{\hat{Q}_h^{\lambda,i}\}_{i,h=1}^{N,H}$.

5.4 Theoretical Analysis for MonoGMFG-PMD with General Function Classes

In this section, we remove the requirement that one is given oracle access to an action-value function and we estimate it in Line 3 of MonoGMFG-PMD using general function classes. We consider the action-value function class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$, where $\mathcal{F}_h \subseteq \{f : \mathcal{S} \times \mathcal{A} \to [0, (H-h+1)(1+\lambda \log|\mathcal{A}|)]\}$ is the class of action-value functions at time $h \in [H]$. Then we estimate the action-value functions using Algorithm 2.

This algorithm mainly involves two steps. The first is involves data collection (Line 3). Here we assign policies to players and collect data from their interactions. We let the sampled N players implement behavior policies $\{\pi_t^{\text{b},i}\}_{i=1}^{N}$, which can be different from $\{\pi_t^{i/N}\}_{i=1}^{N}$. This will not change the aggregate $z_h^{\alpha}(\mu_h^{\mathcal{I}}, W_h)$ for any $\alpha \in \mathcal{I}$, since only a zero-measure set of players change their policies. The second is the action-value function estimation (Lines 7 and 8). The action-value function is selected based on the previous value function, and the value function is updated from the derived estimate. This can be implemented in parallel for all players. We highlight that the estimation analysis cannot leverage results from general non-parametric regression [Wainwright, 2019], since the response variable $\hat{V}_{h+1}^{\lambda,i}$ is *not* independent of $s_{\tau,h+1}^i$ in our setting.

Assumption 5 (Realizability). *For any policy $\pi^{\mathcal{I}} \in \tilde{\Pi}$ and the induced distribution flow $\mu^{\mathcal{I}} \in \tilde{\Delta}$, we have $Q_h^{\lambda,\alpha}(\cdot,\cdot,\pi^{\alpha},\mu^{\mathcal{I}}) \in \mathcal{F}_h$ for $h \in [H]$.*

This assumption ensures that we can find the nominal action-value function in the function class. For a policy $\pi \in \Pi^H$ and a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the operator $(\mathcal{T}_h^{\pi} f)(s,a) = \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[\langle f(s',\cdot), \pi_{h+1}(\cdot|s') \rangle - \lambda R(\pi_{h+1}(\cdot|s'))]$. For a policy $\pi^{\mathcal{I}}$ and the induced distribution flow $\mu^{\mathcal{I}}$, we have $Q_h^{\lambda,\alpha}(s,a,\pi^{\alpha},\mu^{\mathcal{I}}) = r_h(s,a,z_h^{\alpha}) + (\mathcal{T}_h^{\pi^{\alpha}} Q_{h+1}^{\lambda,\alpha})(s,a)$.

Assumption 6 (Completeness). *For any policy $\pi^{\mathcal{I}} \in \tilde{\Pi}$ and the induced distribution flow $\mu^{\mathcal{I}} \in \tilde{\Delta}$, we have that for all $f \in \mathcal{F}_{h+1}$, $r_h(\cdot,\cdot,z_h^{\alpha}(\mu_h^{\mathcal{I}}, W_h)) + (\mathcal{T}_h^{\pi^{\alpha}} f) \in \mathcal{F}_h$ for all $\alpha \in \mathcal{I}$, $h \in [H-1]$.*
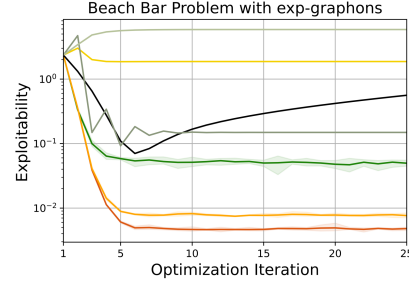
This completeness assumption ensures that the estimates from $\mathcal{F}$ also satisfy the relationship between nominal action-value functions through $\mathcal{T}_h^{\pi^{\alpha}}$. These realizability and completeness assumptions are widely adopted in the off-policy evaluation and offline reinforcement learning literature [Uehara et al., 2022, Xie et al., 2021b].

Assumption 7. *The reward functions $\{r_h\}_{h=1}^{H}$ are Lipschitz in z, i.e., $|r_h(s,a,z) - r_h(s,a,z')| \leq L_r \|z - z'\|_1$ for all $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$. The graphons W_h for $h \in [H]$ are Lipschitz continuous functions, i.e., there exists a constant $L = L_W > 0$ (depending only on $W = \{W_h\}_{h=1}^{H}$) such that $|W_h(\alpha,\beta) - W_h(\alpha',\beta')| \leq L_W(|\alpha - \alpha'| + |\beta - \beta'|)$ for all $\alpha, \alpha', \beta, \beta' \in [0,1]$, and $h \in [H]$.*

The Lipschitzness assumption is common in the GMFG works [Parise and Ozdaglar, 2019, Carmona et al., 2022, Cui and Koeppl, 2021b]. It helps us to approximate the action-value function of a player by that of sampled players. We denote the state distributions of player i induced by policy $\pi_t^{\text{b},i}$ as $\mu_t^{\text{b},i}$. Then we require the behavior policies $\{\pi_t^{\text{b},i}\}_{i=1}^{N}$ to satisfy the following requirements.
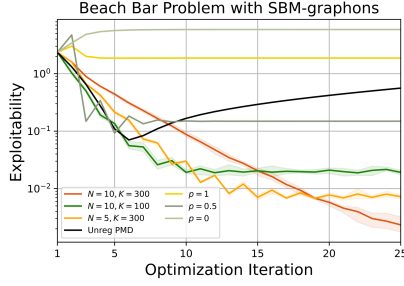
(a) Beach Bar problem with SBM graphons   (b) Beach Bar problem with exp-graphons.

Figure 1: Simulation results for Beach Bar problem with SBM and exp-graphons.

Assumption 8. *For any $t \in [T]$, the behavior policies explore sufficiently. More precisely, for any policy $\pi \in \Pi^H$ and induced distributions $\mu \in \Delta(\mathcal{S})^H$, we have $\sup_{s \in \mathcal{S}, a \in \mathcal{A}} \pi_h(a \,|\, s)/\pi_{t,h}^{\mathrm{b},i}(a \,|\, s) \le C_1$ and $\sup_{s \in \mathcal{S}} \mathrm{d}\mu_h(s)/\mathrm{d}\mu_{t,h}^{\mathrm{b},i}(s) \le C_2$ for $h \in [H]$ and $i \in [N]$ where $C_1, C_2 > 0$ are constants.*

This assumption guarantees that the behavior policies explore the actions that may be adopted by $\pi_t^{\mathcal{I}}$ and $\pi^{*,\mathcal{I}}$. Such an assumption is widely adopted in offline reinforcement learning and off-policy evaluation works [Uehara et al., 2022, Xie et al., 2021b].

Theorem 4. *Assume that the GMFG is weakly monotone and that Assumptions 5, 6, 7, and 8 hold. Let $\eta_t = \eta = O(T^{-1/2})$ and $\beta_t = \beta = O(T^{-1})$ in Algorithm 1 (MonoGMFG-PMD). Then with probability at least $1 - \delta$, Algorithms 1 and 2 yield*

$$D\left(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\mathcal{I}}\right) = O\left(\frac{\lambda \log^2 T}{\sqrt{T}} + C_1 C_2 \frac{H^{3/2}B_H^2}{\lambda\sqrt{K}}\log\frac{TNH \cdot \mathcal{N}_\infty(5B_H/K, \mathcal{F}_{[H]})}{\delta} + \frac{H \log T}{N}\right),$$

where $B_H = H(1 + \lambda \log |\mathcal{A}|)$, and $\mathcal{N}_\infty(5B_H/K, \mathcal{F}_{[H]}) = \max_{h \in [H]} \mathcal{N}_\infty(5B_H/K, \mathcal{F}_h)$ is the ℓ_∞-covering number of the function class.

The error in Theorem 4 consists of both the optimization and estimation errors. The optimization error corresponds to the first term, which also appears in Theorem 3. The estimation error consists of the generalization error and the approximation error, in the second and third terms respectively. When the function class $\mathcal{F}$ is finite, this term scales as $O(K^{-1/2})$, which originates from the fact that we estimate the action-value function from the empirical error instead of its population counterpart. The approximation error scales as $O(N^{-1})$. This term originates from the fact that the action-value function of player α is approximated by that of the sampled player near α. To learn a policy that is at most $\varepsilon > 0$ far from the NE, we can set $T = \tilde{O}(\varepsilon^{-2})$, $K = \tilde{O}(\varepsilon^{-2})$, and $N = \tilde{O}(\varepsilon^{-1})$, which in total results in $TK = \tilde{O}(\varepsilon^{-4})$ episodes of online plays.

6 Experiments

In this section, we conduct experiments to corroborate our theoretical findings. We run different algorithms on the Beach Bar problem Perrin et al. [2020], Fabian et al. [2022]. The underlying graphons are set to Stochastic Block Model (SBM) and exp-graphons. The details of experiments are deferred to Appendix B. Since the NEs of the games are not available, we adopt the *exploitability* to measure the proximity between a policy and the NE. For a policy $\pi^{\mathcal{I}}$ and its induced distribution flow $\mu^{\mathcal{I}}$, the exploitability for the λ-regularized GMFG is defined as

$$\mathrm{Exploit}(\pi^{\mathcal{I}}) = \int_0^1 \max_{\tilde{\pi} \in \Pi^H} J^{\lambda,\alpha}(\tilde{\pi}, \mu^{\mathcal{I}}) - J^{\lambda,\alpha}(\pi^\alpha, \mu^{\mathcal{I}}) \mathrm{d}\alpha.$$

First, the experimental results demonstrate the necessity of modelling the heterogeneity of agents. Figure 1 demonstrates the performance degradation of approximating GMFG by MFG. Here we let the agents play in the GMFG with constant graphons $W_h(\alpha, \beta) = p$ for $p \in \{0, 0.5, 1\}$. The agents have oracle access to the action-value function. We observe that this approximation results in gross errors for learning the NEs of GMFGs with non-constant graphons.

Second, the experiments show that the algorithms designed for unregularized GMFG cannot learn the NE of regularized GMFG. We implement the discrete-time version of the algorithm in Fabian et al.

[2022]; results marked "Unreg PMD" show that the exploitability first decreases and then increases. In line with the discussion in Section 5.2, this originates from keeping too much gradient knowledge in previous iterates $\pi_t^{\mathcal{I}}$. The gradient of the policy is largely correct in the several initial iterations, but a large amount of past knowledge results in it deviating in later iterations, since the past knowledge accumulates. In contrast, our algorithm *discounts* the past knowledge as $(\pi_t^{\mathcal{I}})^{1-\lambda\eta_t}$.

Finally, the results indicate the influence of action-value function estimation. In the experiments, we run our algorithm when $N = 5, K = 300$, $N = 10, K = 100$, and $N = 10, K = 300$. Figure 1 shows that the algorithm with $N = 10, K = 300$ can achieve a smaller error than the algorithms both with $N = 10, K = 100$ and $N = 5, K = 300$. This is in agreement with Theorem 4.

7   Conclusion

In this paper, we focused on two fundamental problems of λ-regularized GMFG. Firstly, we established the existence of NE. This result greatly weakened the conditions in the previous works. Secondly, the provably efficient NE learning algorithms were proposed and analyzed in the weakly monotone GMFG motivated by Lasry and Lions [2007]. The convergence rate of MonoGMFG-PMD features the first performance guarantee of discrete-time algorithm without extra conditions in monotone GMFGs. We leave the lower bound of this problem to the future works.

Acknowledgments and Disclosure of Funding

References

Y. Achdou, P. Cardaliaguet, F. Delarue, A. Porretta, F. Santambrogio, Y. Achdou, and M. Laurière. Mean field games and applications: Numerical aspects. *Mean Field Games: Cetraro, Italy 2019*, pages 249–307, 2020.

B. Anahtarci, C. D. Kariksiz, and N. Saldi. Fitted q-learning in mean-field games. *arXiv preprint arXiv:1912.13309*, 2019.

B. Anahtarci, C. D. Kariksiz, and N. Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, pages 1–29, 2022.

A. Aurell, R. Carmona, G. Dayanıklı, and M. Laurière. Finite state graphon games with applications to epidemics. *Dynamic Games and Applications*, 12(1):49–81, 2022a.

A. Aurell, R. Carmona, and M. Lauriere. Stochastic graphon games: Ii. the linear-quadratic case. *Applied Mathematics & Optimization*, 85(3):1–33, 2022b.

D. Bertsekas and S. E. Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

P. E. Caines and M. Huang. Graphon mean field games and the gmfg equations: ε-nash equilibria. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 286–292. IEEE, 2019.

R. Carmona, D. B. Cooney, C. V. Graves, and M. Lauriere. Stochastic graphon games: I. the static case. *Mathematics of Operations Research*, 47(1):750–778, 2022.

S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

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

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

Kai Cui and Heinz Koeppl. Learning graphon mean field games and approximate nash equilibria. In *International Conference on Learning Representations*, 2021b.

C. Fabian, . Cui, and H. Koeppl. Learning sparse graphon mean field games. *arXiv preprint arXiv:2209.03880*, 2022.

M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

M. Geist, J. Pérolat, M. Laurière, R. Elie, S. Perrin, O. Bachem, R. Munos, and O. Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

A. H. Guide. *Infinite dimensional analysis*. Springer, 2006.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

K. Hinderer. Foundations of non-stationary dynamic programming with discrete time parameter, 1970.

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

H.-J. Langen. Convergence of dynamic programming models. *Mathematics of Operations Research*, 6(4):493–512, 1981.

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

R. Lowe, Y. I. Wu, A. Tamar, J. Harb, OpenAI Pieter A., and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Francesca Parise and Asuman Ozdaglar. Graphon games. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 457–458, 2019.

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

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

L. Shani, Y. Efroni, and S. Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020.

E. Sonu, Y. Chen, and P. Doshi. Decision-theoretic planning under anonymity in agent populations. *Journal of Artificial Intelligence Research*, 59:725–770, 2017.

M. Uehara, C. Shi, and N. Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

D. Vasal, R. K. Mishra, and S. Vishwanath. Master equation of discrete time graphon mean field games and teams. *arXiv preprint arXiv:2001.05633*, 2020.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

L. Wang, Z. Yang, and Z. Wang. Breaking the curse of many agents: Provable mean embedding Q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pages 10092–10103. PMLR, 2020.

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

T. Xie, C. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021b.

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

M. A. uz Zaman, A. Koppel, S. Bhatt, and T. Basar. Oracle-free reinforcement learning in mean-field games along a single sample path. *arXiv preprint arXiv:2208.11639*, 2022.

Appendix for
"Learning Regularized Monotone Graphon Mean-Field Games"

A Detailed Explanations of Table 1

We first explain Table 1 column by column. The first column lists the conditions required by each work. Although the detailed statements of these conditions are usually different, these conditions can be largely categorized into contraction conditions and monotone conditions. Here 'potential' means the extra potential reward structure required in Geist et al. [2021].

'No population manipulation' means that during the learning process, the distribution flow is indeed induced by the current policy. For example, Xie et al. [2021a] and Perrin et al. [2020] mix the current distribution flow with the previous ones to form the distribution flow required by the next step. In contrast, the distribution flows required in algorithms in Perolat et al. [2021], Fabian et al. [2022] and our work are those induced by the policies in each step.

'Online playing' means that the algorithms can be implemented with the data collected from the online playing of agents. In general, the algorithms that do not require population manipulations can be implemented by letting agents play their policies in the online game. Thus, these algorithms admit online playing. In contrast, Anahtarci et al. [2022], Perrin et al. [2020] and Geist et al. [2021] need to solve the optimal policy on specific distribution flows. Thus, they need the access to a simulator for this purpose.

'Heterogeneity' means the modeling of the heterogeneity among agents. The works for MFG only consider homogeneous agents, and thus cannot model the heterogeneity.

'Discrete-time algorithm' means the provably efficient discrete-time algorithms here. Although some discrete-time algorithms are provided in Perrin et al. [2020],Perolat et al. [2021] and Fabian et al. [2022], neither the consistency nor the convergence rate is provided therein.

'Convergence rate' in the final column refers to the convergence rate of both discrete-time algorithms and continuous-time algorithms. Perrin et al. [2020] provides the convergence rate for their continuous-time algorithm, and other works with 'Yes' all provide the convergence rate for the discrete-time algorithms.

In summary, our work provides the first provably efficient discrete-time algorithm in the monotone GMFG without any extra conditions. This result deepen our understanding of the monotone GMFGs, as a complementary setting of contractive GMFGs.

B Experiment Details

We adopt the Beach Bar problem as our GMFG. This problem is initially proposed in Perolat et al. [2021], Perrin et al. [2020] for MFG and modified by Fabian et al. [2022] to GMFG. In the Beach Bar problem, Agents can move their towels between locations and try to be close to the bar but also avoid crowded areas and neighbors in an underlying network. The state space $\mathcal{S}$ is $\{1, 2, \cdots, |\mathcal{S}|\}$, and we set $|\mathcal{S}| = 10$ in our experiments. The bar is located at $B = |\mathcal{S}|/2$. The action space is $\mathcal{A} = \{-1, 0, 1\}$, which indicates the movement of the towel. The transition kernel is $s_{t+1}^\alpha = s_t^\alpha + a_t^\alpha + \varepsilon_t^\alpha$, where ε_t^α is the noise that takes $+1$ or -1 with probability $1/2$. The reward function is defined as

$$r_t(s_t^\alpha, a_t^\alpha, z_t^\alpha) = \frac{2}{|\mathcal{S}|}|B - s_t^\alpha| + \frac{2}{|\mathcal{S}|}|a_t^\alpha| - 8z_t^\alpha \text{ for all } t \in [H].$$

In our experiments, we regualrize this reward function with $\lambda = 1$. The underlying graphons in our experiments are SBM and exp-graphons. The exp-graphon is defined as

$$W_\theta^{\exp}(\alpha, \beta) = \frac{2\exp(\theta \cdot \alpha\beta)}{1 + \exp(\theta \cdot \alpha\beta)} - 1,$$

where $\theta > 0$ is the parameter. In our simulation, we set $\theta = 3$. The SBM in our experiments has two communities with 70% and 30% population respectively. The inter-community rate is 0.3, and the intra-community rate is 0.9. In our experiments, we adopt the exploitability to measure the closeness between a policy and the NE.

Since the Beach Bar problem only involves the finite state and action spaces, our algorithms take the function class $\mathcal{F}_h = \{f : \mathcal{S} \times \mathcal{A} \to [0, H(1 + \lambda \log |\mathcal{A}|)]\}$ for all $h \in [H]$.

Figure 1 is generated from five Monte-Carlo implementations for each algorithm. The error bar in the figure indicates the maximal and the minimal error in the Monte-Carlo. To simulate cases with constant graphons, we implement mirror descent algorithm with the nominal action-value functions, which are directly calculated from the ground-truth transition kernels and reward functions. Thus, there is no error bar for thm. To simulate the policy mirror descent for unregularized GMFG, we directly use the code of Fabian et al. [2022], and the action-value functions are also calculated from the ground-truth model. Thus, there is no error bar for it, either. Our simulations run on a single Intel(R) Xeon(R) CPU E5-2697 v4 @ 2.30 GHz, and the experiments take about two days,

C Proof of Theorem 1

Proof of Theorem 1. We prove the existence of NE by three steps:

- We construct a λ-regularized MFG based on the λ-regularized GMFG.

- We show that we can construct an NE of the λ-regularized GMFG from an NE of the constructed λ-regularized MFG.

- We show that the constructed λ-regularized MFG has NE under Assumptions **??** and 3.

Step 1: Construction of a λ-regularized MFG

The state and action spaces of the λ-regularized MFG is $\bar{\mathcal{S}} = \mathcal{S} \times \mathcal{I}$ and $\mathcal{A}$ respectively, where $\mathcal{I} = [0, 1]$. Here, we treat the positions of players as a state in MFG. The state of the player is denoted as $\bar{s}_h = (s_h, \alpha_h) \in \bar{\mathcal{S}}$, and we denote the distribution of the state at time h as $\mathcal{L}_h = \mathcal{L}(\bar{s}_h) = \mathcal{L}(s_h, \alpha_h)$, which is the law of state $\bar{s}_t$.

At time h, the transition kernel of such MFG is $\bar{P}_h : \bar{\mathcal{S}} \times \mathcal{A} \times \Delta(\bar{\mathcal{S}}) \to \Delta(\bar{\mathcal{S}})$. To specify $\bar{P}_h$, we first define a function of α_h and $\mathcal{L}_h$ as $f(\alpha_h, \mathcal{L}_h, W_h) : \mathcal{I} \times \Delta(\mathcal{S} \times \mathcal{I}) \times \mathcal{W} \to \mathcal{M}(\mathcal{S})$, whose output is a measure supported on $\mathcal{S}$, i.e.,

$$\big[f(\alpha_h, \mathcal{L}_h, W_h) \big](\cdot) = \int_0^1 W_h(\alpha_h, \beta) \mathcal{L}_h(\cdot, \beta) \mathrm{d}\beta. \tag{C.1}$$

Eqn. (C.1) enables us to define the transition kernel $\bar{P}_h$ as

$$\bar{P}_h\big(\bar{s}_{h+1} \,|\, \bar{s}_h, a_h, \mathcal{L}_h\big) = \delta_{\alpha_{h+1} = \alpha_h} \cdot P_h\big(s_{h+1} \,|\, s_h, a_h, f(\alpha_h, \mathcal{L}_h, W_h)\big)$$

$$= \delta_{\alpha_{h+1} = \alpha_h} \cdot P_h\Big(s_{h+1} \,|\, s_h, a_h, \int_0^1 W_h(\alpha_h, \beta) \mathcal{L}_h(\cdot, \beta) \mathrm{d}\beta\Big), \tag{C.2}$$

where $\delta_{\alpha_{h+1} = \alpha_h}$ is the Dirac's delta function at α_h, and P_h is the transition kernel of the GMFG. The reward function of the MFG can be similarly defined as

$$\bar{r}_h(\bar{s}_h, a_h, \mathcal{L}_h) = r_h\Big(s_h, a_h, \int_0^1 W_h(\alpha_h, \beta) \mathcal{L}_h(\cdot, \beta) \mathrm{d}\beta\Big). \tag{C.3}$$

The initial state distribution of the MFG is specified as $\mathcal{L}_1 = \mu_1 \times \mathrm{Unif}([0, 1])$. The value functions of the λ-regularized MFG are defined as

$$\bar{V}_h^\lambda\big((s, \alpha), \pi, \mathcal{L}\big) = \mathbb{E}^\pi\left[\sum_{t=h}^H \bar{r}_h(\bar{s}_t, a_t, \mathcal{L}_t) - \lambda \log \pi_t(a_t \,|\, s_t, \alpha_t) \,\bigg|\, s_h = s, \alpha_h = \alpha \right],$$

where the expectation $\mathbb{E}^\pi$ is taken with respect to the MDP $a_t \sim \pi_t(\cdot \,|\, s_t, \alpha_t)$ and $\bar{s}_{t+1} \sim \bar{P}_t(\cdot \,|\, \bar{s}_t, a_t, \mathcal{L}_t)$ for $t \in [H]$. Then the cumulative reward function is defined as

$$\bar{J}^\lambda(\pi, \mathcal{L}) = \mathbb{E}_{\mathcal{L}_1}\big[\bar{V}_1^\lambda(s, \alpha, \pi, \mathcal{L}) \big]. \tag{C.4}$$

Step 2: Construction of the NE of the λ-regularized GMFG from the NE of the λ-regularized MFG

In this step, we assume that the λ-regularized MFG defined in Eqn. (C.2) and (C.3) admits an NE $(\tilde{\pi}, \tilde{\mathcal{L}})$, which is defined in Definition 2, replacing the discounted reward function therein by the reward defined in Eqn. (C.4). We will construct a policy and distribution flow pair $(\pi^{\mathcal{I}}, \mu^{\mathcal{I}})$ of the λ-regularized GMFG from $(\tilde{\pi}, \tilde{\mathcal{L}})$ and show that $(\pi^{\mathcal{I}}, \mu^{\mathcal{I}})$ is indeed an NE of the λ-regularized GMFG.

We construct the policy and distribution flow pair as $\pi_h^\alpha(\cdot \mid s) = \tilde{\pi}_h(\cdot \mid s, \alpha)$ and $\mu_h^\alpha(\cdot) = \tilde{\mathcal{L}}_h(\cdot, \alpha)$ for all $h \in [H]$, $s \in \mathcal{S}$, and $\alpha \in \mathcal{I}$. To prove that $(\pi^{\mathcal{I}}, \mu^{\mathcal{I}})$ is an NE of GMFG, we need to show: (i) $\mu^{\mathcal{I}}$ is induced by the policy $\pi^{\mathcal{I}}$, i.e., $\Gamma_2(\pi^{\mathcal{I}}, W) = \mu^{\mathcal{I}}$. (ii) $\pi^{\mathcal{I}}$ is the optimal policy given $\mu^{\mathcal{I}}$ for all the players.

We use induction to prove (i). Define $\tilde{\mu}^{\mathcal{I}} = \Gamma_2(\pi^{\mathcal{I}}, W)$. We will show that $\tilde{\mu}^{\mathcal{I}} = \mu^{\mathcal{I}}$. For $h = 1$, we have $\mu_1'^{,\alpha} = \mu_1 = \mathcal{L}_1(\cdot, \alpha) = \mu_1^\alpha$ for all $\alpha \in \mathcal{I}$. Assume that $\mu_h'^{,\alpha} = \mu_h^\alpha$ for all $\alpha \in \mathcal{I}$, for time $h + 1$ and any $s' \in \mathcal{S}$ and $\alpha \in \mathcal{I}$, we have

$$
\begin{aligned}
\mu_{h+1}'^{,\alpha}(s') &= \int_{\mathcal{S}} \int_{\mathcal{A}} \mu_h'^{,\alpha}(s) \pi_h^\alpha(a \mid s) P_h\left(s' \mid s, a, \int_0^1 W_h(\alpha, \beta) \mu_h'^{,\alpha} \mathrm{d}\beta\right) \mathrm{d}a \mathrm{d}s \\
&= \int_{\mathcal{S}} \int_{\mathcal{A}} \int_0^1 \tilde{\mathcal{L}}_h(s, \alpha') \tilde{\pi}_h(a \mid s, \alpha') P_h\left(s' \mid s, a, \int_0^1 W_h(\alpha', \beta) \tilde{\mathcal{L}}_h(\cdot, \beta) \mathrm{d}\beta\right) \delta_{\alpha=\alpha'} \mathrm{d}\alpha' \mathrm{d}a \mathrm{d}s \\
&= \tilde{\mathcal{L}}_{h+1}(s', \alpha),
\end{aligned}
$$

where the first equation follows from the definition of $\tilde{\mu}^{\mathcal{I}}$, the second equation follows from the fact that $\mu_h'^{,\alpha} = \mu_h^\alpha$ and $\mu_h^\alpha(\cdot) = \tilde{\mathcal{L}}_h(\cdot, \alpha)$, and the last equation follows from the definition of $\tilde{\mathcal{L}}$. Then (i) results from the fact that $\mu_{h+1}^\alpha(\cdot) = \tilde{\mathcal{L}}_{h+1}(\cdot, \alpha) = \mu_{h+1}'^{,\alpha}(\cdot)$.

To prove (ii), we compare the MDPs given $\tilde{\mathcal{L}}$ and $\mu^{\mathcal{I}}$ in MFG and GMFG. The MDP for the player α in GMFG is specified by the transition kernel $s_{h+1}^\alpha \sim P_h(\cdot \mid s_h^\alpha, a_h^\alpha, \int_0^1 W_h(\alpha, \beta) \mu_h^\beta \mathrm{d}\beta)$, and the reward function $r_h(s_h^\alpha, a_h^\alpha, \int_0^1 W_h(\alpha, \beta) \mu_h^\beta \mathrm{d}\beta)$. We want to prove that $V_1^{\lambda,\alpha}(s, \pi^\alpha, \mu^{\mathcal{I}}) \geq V_1^{\lambda,\alpha}(s, \bar{\pi}^\alpha, \mu^{\mathcal{I}})$ for all $s \in \mathcal{S}$, and $\bar{\pi}^\alpha \in \Pi$.

Since $\tilde{\pi}$ is optimal with respect to $\tilde{\mathcal{L}}$, we have $\bar{V}_1^\lambda(s, \alpha, \tilde{\pi}, \tilde{L}) \geq \bar{V}_1^\lambda(s, \alpha, \bar{\pi}, \tilde{L})$ for all $s \in \mathcal{S}$, $\alpha \in \mathcal{I}$, and policy $\bar{\pi}$. Given $s_1 = s$ and $\alpha_1 = \alpha$, the MDP in MFG is specified by the transition kernel $s_{h+1} \sim P_h(s_{h+1} \mid s_h, a_h, \int_0^1 W_h(\alpha, \beta) \tilde{\mathcal{L}}_t(\cdot, \beta) \mathrm{d}\beta)$, $\alpha_{h+1} = \alpha$, and the reward function $r_h(s_h^\alpha, a_h^\alpha, \int_0^1 W_h(\alpha, \beta) \tilde{\mathcal{L}}_t(\cdot, \beta) \mathrm{d}\beta)$. We note that these two MDPs are the same, and $V_1^{\lambda,\alpha}(s, \pi^\alpha, \mu^{\mathcal{I}}) = \bar{V}_1^\lambda(s, \alpha, \tilde{\pi}, \tilde{L})$. This proves the claim (ii).

Step 3: Prove the existence of NE in the constructed MFG under Assumptions 2 and 3.

In order to prove the existence of NE in the constructed MFG, we only need to verify Assumption 4 in Theorem 2.

We first verify Assumption 4 (1) and (2) hold. Our reward functions $\bar{r}_h$ are bounded, $\mathcal{A}$ is finite, and the state space $\mathcal{S} \times \mathcal{I}$ is compact.

For Assumption 4 (3) and (4), we only need to verify that the reward function in Eqn. (C.3) is continuous and the transition kernel in Eqn. (C.2) is continuous with respect to total variation. Since r_h is continuous, we only need to prove that $\int_0^1 W_h(\alpha_h, \beta) \mathcal{L}_h(\cdot, \beta) \mathrm{d}\beta$ is continuous for the continuity of $\bar{r}_h$. In the following, we make use of the fact that the convergence in total variation implies the weak convergence.

Given two sequences $\{\alpha_n\}$ and $\{\mathcal{L}_n\}$ such that $\alpha_n \to \alpha$ and $\mathcal{L}_n$ converges to $\mathcal{L}$ in total variation, we have

$$
\begin{aligned}
&\int_{\mathcal{S}} \left| \int_0^1 W_h(\alpha_n, \beta) \mathcal{L}_n(s, \beta) \mathrm{d}\beta - \int_0^1 W_h(\alpha, \beta) \mathcal{L}(s, \beta) \mathrm{d}\beta \right| \mathrm{d}s \\
&\leq \int_{\mathcal{S}} \left| \int_0^1 W_h(\alpha_n, \beta) \mathcal{L}_n(s, \beta) \mathrm{d}\beta - \int_0^1 W_h(\alpha, \beta) \mathcal{L}_n(s, \beta) \mathrm{d}\beta \right| \mathrm{d}s \\
&\quad + \int_{\mathcal{S}} \left| \int_0^1 W_h(\alpha, \beta) \mathcal{L}_n(s, \beta) \mathrm{d}\beta - \int_0^1 W_h(\alpha, \beta) \mathcal{L}(s, \beta) \mathrm{d}\beta \right| \mathrm{d}s
\end{aligned}
$$

15

$$\leq \int_{\mathcal{S}} \left| \int_0^1 W_h(\alpha_n, \beta) \mathcal{L}_n(s, \beta) \mathrm{d}\beta - \int_0^1 W_h(\alpha, \beta) \mathcal{L}_n(s, \beta) \mathrm{d}\beta \right| \mathrm{d}s$$

$$+ \int_{\mathcal{S}} \int_0^1 |\mathcal{L}_n(s, \beta) - \mathcal{L}(s, \beta)| \mathrm{d}\beta \mathrm{d}s, \tag{C.5}$$

where the first inequality follows from the triangle inequality. Since $\alpha_n \to \alpha$ and the uniform continuity of graphons, the first term in the right-hand side of inequality (C.5) tends to 0. Since $\mathcal{L}_n$ converges to $\mathcal{L}$ in total variation, the second term in the right-hand side of inequality (C.5) tends to 0. Thus, the reward function is continuous, which verifies Assumption 4 (3). For any $g \in C_b(\mathcal{S} \times \mathcal{I})$, given four sequences $\{s_n\}$, $\{\alpha_n\}$, $\{a_n\}$, and $\{L_n\}$ such that $s_n \to s$, $\alpha_n \to \alpha$, $a_n \to a$, and L_n weakly converges to L, we have

$$\left| \int_0^1 \int_{\mathcal{S}} g(s', \alpha') \big[\bar{P}_h(s', \alpha' \mid \bar{s}_h, a_h, \mathcal{L}_h) - \bar{P}_h(s', \alpha' \mid \bar{s}, a, \mathcal{L}) \big] \mathrm{d}s' \mathrm{d}\alpha' \right|$$

$$= \left| \int_{\mathcal{S}} g(s', \alpha_n) P_h\Big(s' \,\Big|\, s_n, a_n, z_h^{\alpha_n}(\mathcal{L}_n, W_h) \Big) \mathrm{d}s' - \int_{\mathcal{S}} g(s', \alpha) P_h\Big(s' \,\Big|\, s, a, z_h^{\alpha}(\mathcal{L}, W_h) \Big) \mathrm{d}s' \right|$$

$$\leq \left| \int_{\mathcal{S}} g(s', \alpha_n) P_h\Big(s' \,\Big|\, s_n, a_n, z_h^{\alpha_n}(\mathcal{L}_n, W_h) \Big) \mathrm{d}s' - \int_{\mathcal{S}} g(s', \alpha) P_h\Big(s' \,\Big|\, s, a, z_h^{\alpha_n}(\mathcal{L}_n, W_h) \Big) \mathrm{d}s' \right|$$

$$+ \left| \int_{\mathcal{S}} g(s', \alpha) P_h\Big(s' \,\Big|\, s, a, z_h^{\alpha_n}(\mathcal{L}_n, W_h) \Big) \mathrm{d}s' - \int_{\mathcal{S}} g(s', \alpha) P_h\Big(s' \,\Big|\, s, a, z_h^{\alpha}(\mathcal{L}, W_h) \Big) \mathrm{d}s' \right|, \tag{C.6}$$

where $z_h^{\alpha_n}(\mathcal{L}_n, W_h) = \int_0^1 W_h(\alpha_n, \beta) \mathcal{L}_n(\cdot, \beta) \mathrm{d}\beta$, $z_h^{\alpha}(\mathcal{L}, W_h) = \int_0^1 W_h(\alpha, \beta) \mathcal{L}(\cdot, \beta) \mathrm{d}\beta$, the equation follows from Eqn. (C.2), and the inequality follows from the triangle inequality. Since $g \in C_b(\mathcal{S} \times \mathcal{I})$ is a continuous function on a compact set and $\alpha_n \to \alpha$, the first term in the right-hand side of inequality (C.6) tends to 0. Since inequality (C.5) proves that $z_h^{\alpha_n}(\mathcal{L}_n, W_h)$ converges to $z_h^{\alpha}(\mathcal{L}, W_h)$ in ℓ_1 and $g(\cdot, \alpha) \in C_b(\mathcal{S})$, the second term in the right-hand side of inequality (C.6) tends to 0. Thus, the transition kernel is continuous, which verifies Assumption 4. It concludes the verification of Assumption 4 in Theorem 2. Thus, we conclude the proof of Theorem 1. $\square$

D   Proof of Theorem 2

Proof of Theorem 2. We prove the existence of NE by two steps:

- We construct an operator Γ that is defined for the state-action distribution flow and show that we can construct the NE from the fixed point of this operator.

- We show that the fixed point set of the operator Γ is not empty.

Step 1: Construction of an operator Γ.

Without the loss of generality, we assume that $r_h : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S}) \to [0, 1]$ for $h \in [H]$. Define constants $L_h = (H - h + 1)(1 + \lambda \log |\mathcal{A}|)$ for $h \in [H]$. Given the continuous functions set $C(\mathcal{S})$, we define the L_h−bounded continuous function set $C_h = \{ f \in C(\mathcal{S}) \,|\, \sup_{s \in \mathcal{S}} |f(s)| \leq L_h \}$ and the product of them $\mathcal{C} = \prod_{h=1}^H C_h(\mathcal{S})$. Given a constant $0 < \sigma < 1$, we equip $\mathcal{C}$ with the metric $\rho(u, v) = \sum_{h=1}^H \sigma^{-h} \|u_h - v_h\|_\infty$ for any $u, v \in \mathcal{C}$ and $\|f\|_\infty = \sup_{s \in \mathcal{S}} |f(s)|$. Then $(\mathcal{C}, \rho)$ is complete.

We define the state-action distribution flow set as $\Xi = \prod_{h=1}^H \Delta(\mathcal{S} \times \mathcal{A})$. For ease of notation, we denote the marginalization of any $\nu \in \Xi$ on $\mathcal{S}$ as $\bar{\nu}_h(s) = \sum_{a \in \mathcal{A}} \nu_h(s, a)$.

For any $\nu \in \Xi$, define an operator T_h^ν acting on $\mathcal{S} \to \mathbb{R}$ as

$$T_h^\nu u(s) = \sup_{p \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) \bar{r}_h(s, a, \bar{\nu}_h) - \lambda R(p) + \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} p(a) \bar{P}_h(s' \mid s, a, \bar{\nu}_h) u(s') \mathrm{d}s' \text{ for } h \in [H-1],$$

$$T_H^\nu u(s) = \sup_{p \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) \bar{r}_H(s, a, \bar{\nu}_H) - \lambda R(p),$$

where $R(\cdot)$ is the negative entropy. When $\lambda = 0$, the supremum is taken with respect to the action $a \in \mathcal{A}$, and the following proposition can be similarly built for $\lambda = 0$.

Proposition 3. *Let $\nu \in \Xi$ be arbitrary distribution flow. For all $h \in [H]$, T_h^ν maps $C_{h+1}(\mathcal{S})$ into $C_h(\mathcal{S})$. In addition, for any $u, v \in C_{h+1}$, we have $\|T_h^\nu u - T_h^\nu v\|_\infty \leq \|u - v\|_\infty$.*

Proof of Proposition 3. See Appendix I.1. □

We then define an operator $T^\nu : \mathcal{C} \to \mathcal{C}$ as $(T^\nu u)_h = T_h^\nu u_{h+1}$ for all $h \in [H]$. We then have

$$\rho(T^\nu u, T^\nu v) = \sum_{h=1}^H \sigma^{-h} \|T_h^\nu u_{h+1} - T_h^\nu v_{h+1}\|_\infty \leq \sum_{h=1}^H \sigma^{-h} \|u_{h+1} - v_{h+1}\|_\infty \leq \sigma \rho(u, v), \text{ (D.1)}$$

where the first inequality results from Proposition 3. Thus, T^ν is a contraction on $(\mathcal{C}, \rho)$, and it has an unique fixed point. For any state-action distribution $\nu \in \Xi$, we use $\bar{V}_h^{\lambda,\nu} : \mathcal{S} \to \mathbb{R}$ for $h \in [H]$ to denote the value functions of the optimal policy in the λ-regularized MDP induced by ν as $\bar{r}_h(s, a, \bar{\nu}_h)$ and $\bar{P}_h(\cdot \mid s, a, \bar{\nu}_h)$ for $h \in [H]$. Then the theory of Markov process shows that [Hinderer, 1970, Theorem 14.1, Theorem 17.1]

Proposition 4. *For any $\nu \in \Xi$, $\bar{V}^{\lambda,\nu} = (\bar{V}_h^{\lambda,\nu})_{h=1}^H$ is the unique fixed point of T^ν. A policy $\pi \in \Pi^H$ is optimal if and only if the following equation holds for any $h \in [H]$ and $\mu_h^+ - a.s.$ state $s \in \mathcal{S}$, where μ^+ is the distribution of states when implementing π on the MDP induced by $\bar{\nu}$.*

$$\sum_{a \in \mathcal{A}} p(a) \bar{r}_h(s, a, \bar{\nu}_h) - \lambda R(p) + \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} p(a) \bar{P}_h(s' \mid s, a, \bar{\nu}_h) \bar{V}_{h+1}^{\lambda,\nu}(s') \mathrm{d}s' = T_h^\nu \bar{V}_{h+1}^{\lambda,\nu}(s).$$

For any $\nu \in \Xi$, we define the sets

$$A(\nu) = \left\{ \xi \in \Xi \,\middle|\, \bar{\xi}_1 = \mu_1, \, \bar{\xi}_{h+1}(\cdot) = \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \bar{P}_h(\cdot \mid s, a, \bar{\nu}_h) \nu_h(\mathrm{d}s, a) \text{ for all } h \in [H-1] \right\},$$

$$B(\nu) = \left\{ \xi \in \Xi \,\middle|\, \text{For all } h \in [H], \sum_{a \in \mathcal{A}} \xi_h(a \mid s) \bar{r}_h(s, a, \bar{\nu}_h) - \lambda R\big(\xi_h(\cdot \mid s)\big) \right.$$

$$\left. + \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \xi_h(a \mid s) \bar{P}_h(s' \mid s, a, \bar{\nu}_h) \bar{V}_{h+1}^{\lambda,\nu}(s') \mathrm{d}s' = \bar{V}_h^{\lambda,\nu}(s), \, \bar{\xi}_h\text{-a.s.} \right\},$$

$$\Gamma(\nu) = A(\nu) \cap B(\nu).$$

We note that $\Gamma(\nu) \neq \emptyset$, since set $A(\nu)$ imposes constraints on $\bar{\xi}_h$ while $B(\nu)$ imposes constraints on $\xi_h(\cdot \mid s)$. We say that ν is a fixed point of Γ if $\nu \in \Gamma(\nu)$.

Step 2: The existence of the fixed point of Γ.

Proposition 5. *Suppose that Γ has a fixed point $\nu \in \xi$. Then we construct a policy as: $\pi_h(\cdot \mid s) = \nu_h(\cdot \mid s)$ for all $s \in \mathrm{supp}(\bar{\nu}_h)$ and $h \in [H]$, $\pi_h(\cdot \mid s)$ for $s \notin \mathrm{supp}(\bar{\nu}_h)$ can be arbitrarily defined. Then the pair $(\pi, \bar{\nu})$ is an NE of the λ-regularized MFG.*

Proof of Proposition 5. Since $\nu \in A(\nu)$, the distribution consistency condition in Definition 2 holds. Since $\nu \in B(\nu)$, the policy defined in Proposition 5 satisfies the optimality condition in Proposition 4, which verifies the player rationality condition in Definition 2. □

Proposition 6. *The graph of Γ, i.e., $Gr(\Gamma) = \{(\nu, \xi) \in \Xi \times \Xi \mid \xi \in \Gamma(\nu)\}$ is closed.*

Proof of Proposition 6. See Appendix I.2. □

The existence of the fixed point of operator Γ follows from the Kakutani's Theorem [Guide, 2006, Corollary 17.55]. We note that the existence of the NE is the direct result of Proposition 4. This concludes the proof of Theorem 2.

□

E Proof of Theorem 3

Proof of Theorem 3. According to the policy update procedures Line 5 and Line 6 in Algorithm 1, we have that

$$
D_{\mathrm{kl}}\big(\pi_h^{*,\alpha}(\cdot \mid s_h^\alpha)\|\pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\big)
$$

$$
\leq D_{\mathrm{kl}}\big(\pi_h^{*,\alpha}(\cdot \mid s_h^\alpha)\|\hat{\pi}_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\big) + \beta/(1-\beta)
$$

$$
\leq -\eta\big\langle \hat{Q}_h^{\lambda,\alpha}(s_h^\alpha,\cdot,\pi_t^\alpha,\mu_t^{\mathcal{I}}) - \lambda\log\pi_h^\alpha(\cdot \mid s_h^\alpha), \pi_h^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\big\rangle
$$

$$
+ D_{\mathrm{kl}}\big(\pi_h^{*,\alpha}(\cdot \mid s_h^\alpha)\|\pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\big) + \frac{1}{2}\eta^2\Big(H + \lambda H\log|\mathcal{A}| + \lambda\log\frac{|\mathcal{A}|}{\beta}\Big)^2 + \frac{\beta}{1-\beta}
$$

$$
\leq -\eta\big\langle Q_h^{\lambda,\alpha}(s_h^\alpha,\cdot,\pi_t^\alpha,\mu_t^{\mathcal{I}}) - \lambda\log\pi_h^\alpha(\cdot \mid s_h^\alpha), \pi_h^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\big\rangle
$$

$$
+ D_{\mathrm{kl}}\big(\pi_h^{*,\alpha}(\cdot \mid s_h^\alpha)\|\pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\big) + \frac{1}{2}\eta^2\Big(H + \lambda H\log|\mathcal{A}| + \lambda\log\frac{|\mathcal{A}|}{\beta}\Big)^2 + \frac{\beta}{1-\beta} + \eta\varepsilon_h^\alpha,
$$

$$
\tag{E.1}
$$

where the first inequality results from Lemma 10, and the second inequality results from Lemma 12, and the last inequality follows from the definition of ε_h^α, which is defined as the upperbound of

$$
\Big|\big\langle \hat{Q}_h^{\lambda,\alpha}(s_h^\alpha,\cdot,\pi_t^\alpha,\mu_t^{\mathcal{I}}) - Q_h^{\lambda,\alpha}(s_h^\alpha,\cdot,\pi_t^\alpha,\mu_t^{\mathcal{I}}), \pi_h^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\big\rangle\Big| \leq \varepsilon_h^\alpha \text{ for all } t \in [T].
$$

Taking expectation with respect to $\mu_h^{*,\alpha}$ on the both sides of inequality (E.1), we can upper bound the difference between $D(\pi_{t+1}^{\mathcal{I}})$ and $D(\pi_t^{\mathcal{I}})$ as

$$
D(\pi_{t+1}^{\mathcal{I}}) - D(\pi_t^{\mathcal{I}})
$$

$$
= \int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}\Big[D_{\mathrm{kl}}\big(\pi_h^{*,\alpha}(\cdot \mid s_h^\alpha)\|\pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\big) - D_{\mathrm{kl}}\big(\pi_h^{*,\alpha}(\cdot \mid s_h^\alpha)\|\pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\big)\Big]\mathrm{d}\alpha
$$

$$
\leq \eta\int_0^1 J^{\lambda,\alpha}(\pi_t^\alpha,\mu_t^{\mathcal{I}}) - J^{\lambda,\alpha}(\pi^{*,\alpha},\mu_t^{\mathcal{I}})\mathrm{d}\alpha - \lambda\eta\int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}\Big[D_{\mathrm{kl}}\big(\pi_h^{*,\alpha}(\cdot \mid s_h^\alpha)\|\pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\big)\Big]\mathrm{d}\alpha
$$

$$
+ \frac{1}{2}\eta^2 H\Big(H + \lambda H\log|\mathcal{A}| + \lambda\log\frac{|\mathcal{A}|}{\beta}\Big)^2 + \frac{\beta}{1-\beta}H + 2\eta\int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}[\varepsilon_h^\alpha]\mathrm{d}\alpha
$$

$$
\leq -\lambda\eta D(\pi_t^{\mathcal{I}}) + \frac{1}{2}\eta^2 H\Big(H + \lambda H\log|\mathcal{A}| + \lambda\log\frac{|\mathcal{A}|}{\beta}\Big)^2 + \frac{\beta}{1-\beta}H + 2\eta\int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}[\varepsilon_h^\alpha]\mathrm{d}\alpha,
$$

$$
\tag{E.2}
$$

where the first equation results from the definition of $H(\cdot)$, the first inequality results from inequality (E.1) and Lemma 11, and the last inequality results from Proposition 2 and the definition of NE. The inequality (E.2) can be reformulated as

$$
D(\pi_t^{\mathcal{I}}) \leq \frac{1}{\lambda\eta}\big(D(\pi_t^{\mathcal{I}}) - D(\pi_{t+1}^{\mathcal{I}})\big) + \frac{\eta}{2\lambda}H\Big(H + \lambda H\log|\mathcal{A}| + \lambda\log\frac{|\mathcal{A}|}{\beta}\Big)^2 + \frac{\beta H}{(1-\beta)\lambda\eta}
$$

$$
+ \frac{2}{\lambda}\int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}[\varepsilon_h^\alpha]\mathrm{d}\alpha.
$$

Thus, we have that

$$
\frac{1}{T}\sum_{t=1}^T D(\pi_t^{\mathcal{I}}) \leq \frac{1}{T\lambda\eta}D(\pi_1^{\mathcal{I}}) + \frac{\eta}{2\lambda}H\Big(H + \lambda H\log|\mathcal{A}| + \lambda\log\frac{|\mathcal{A}|}{\beta}\Big)^2 + \frac{\beta H}{(1-\beta)\lambda\eta}
$$

$$
+ \frac{2}{\lambda}\int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}[\varepsilon_h^\alpha]\mathrm{d}\alpha.
$$

18

Take $\eta = O(T^{-1/2})$ and $\beta = O(T^{-1})$, then we have $\beta/(1-\beta) = O(T^{-1})$. Thus, we have

$$\frac{1}{T}\sum_{t=1}^{T} D(\pi_t^{\mathcal{I}}) = O\left(\frac{\lambda \log^2 T}{T^{1/2}} + \frac{\int_0^1 \sum_{h=1}^{H} \mathbb{E}_{\mu_h^{*,\alpha}}[\varepsilon_h^\alpha]\mathrm{d}\alpha}{\lambda}\right).$$

The desired result in Theorem 3 follows from the convexity of KL divergence. Thus, we conclude the proof of Theorem 3 by noting that $\varepsilon_h^\alpha = 0$ in this case. $\qquad\square$

F Proof of Theorem 4

Proof of Theorem 4. The proof of Theorem 4 mainly involves two steps:

- Derive the performance guarantee for Algorithm 2.

- Combine the estimation result with the optimization result in Theorem 3.

Step 1: Derive the performance guarantee for Algorithm 2.

Now we focus on estimating the action-value function of ith player with policy $\pi_t^{i/N}$. We first introduce some notations. The nominal action-value and the value functions of the ith player with policy $\pi_t^{i/N}$ and underlying distribution flow $\mu_t^{\mathcal{I}}$ are respectively denoted as $Q_h^{\lambda,i}(s,a) = Q_h^{\lambda,i}(s,a,\pi_t^{i/N},\mu_t^{\mathcal{I}})$ and $V_h^{\lambda,i}(s) = V_h^{\lambda,i}(s,\pi_t^{i/N},\mu_t^{\mathcal{I}})$ for $h \in [H]$. In the following, we use z_h^α to denote the aggregate $z_h^\alpha(\mu_{t,h}^{\mathcal{I}},W_h)$. The distribution of $(s_h^i,a_h^i,r_h^i,s_{h+1}^i)$ induced by the behavior policy $\pi_t^{\mathrm{b},i/N}$ and the distribution flow $\mu_t^{\mathcal{I}}$ as $(s_h^i,a_h^i,r_h^i,s_{h+1}^i) \sim \mu_h^{\mathrm{b},i} \times \pi_h^{\mathrm{b},i} \times \delta_{r_h} \times P_h = \tilde{\rho}_h^{\mathrm{b},i}$, and the marginalization of this distribution on (s_h^i,a_h^i,r_h^i) is denoted as $\rho_h^{\mathrm{b},i}$. For any function $f : \mathcal{S} \to \mathbb{R}$, we define an operator $\mathcal{P}_h$ as $(\mathcal{P}_h f)(s,a) = \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[f(s')]$. Then we adopt recurrence on the time step $h \in [H]$ to derive the estimation performance guarantee.

For time $h = H$, Line 7 in Algorithm 2 simplifies to

$$\hat{Q}_H^{\lambda,i} = \operatorname*{argmin}_{f \in \mathcal{F}_h} \sum_{\tau=1}^{K} \left(f(s_{\tau,h}^i, a_{\tau,h}^i) - r_{\tau,h}^i\right)^2$$

This corresponds to the classical non-parametric regression problem, Theorem 11.4 of Györfi et al. [2002] shows that the performance guarantee can be derived as

$$\mathbb{E}_{\rho_H^{\mathrm{b},i}}\left[\left|Q_H^{\lambda,i}(s,a) - \hat{Q}_H^{\lambda,i}(s,a)\right|^2\right] = O\left(\frac{B_H^4}{K}\log\frac{\mathcal{N}_\infty(5B_H/K,\mathcal{F}_H)}{\delta}\right).$$

For a time step $h \in [H-1]$, we have that

$$\mathbb{E}_{\rho_h^{\mathrm{b},i}}\left[\left|Q_h^{\lambda,i}(s,a) - \hat{Q}_h^{\lambda,i}(s,a)\right|^2\right]$$

$$\leq 2\mathbb{E}_{\rho_h^{\mathrm{b},i}}\left[\left|Q_h^{\lambda,i}(s,a) - r_h(s,a,z_h^{i/N}) - (\mathcal{P}_h \hat{V}_{h+1}^{\lambda,i})(s,a)\right|^2\right]$$

$$+ 2\mathbb{E}_{\rho_h^{\mathrm{b},i}}\left[\left|r_h(s,a,z_h^{i/N}) + (\mathcal{P}_h \hat{V}_{h+1}^{\lambda,i})(s,a) - \hat{Q}_h^{\lambda,i}(s,a)\right|^2\right]. \tag{F.1}$$

For the first term in the right-hand side of inequality (F.1), we have

$$\mathbb{E}_{\rho_h^{\mathrm{b},i}}\left[\left|Q_h^{\lambda,i}(s,a) - r_h(s,a,z_h^{i/N}) - (\mathcal{P}_h \hat{V}_{h+1}^{\lambda,i})(s,a)\right|^2\right]$$

$$= \mathbb{E}_{\rho_h^{\mathrm{b},i}}\left[\left|r_h(s,a,z_h^{i/N}) + (\mathcal{T}_h^{\pi_t^{i/N}} Q_{h+1}^{\lambda,i})(s,a) - r_h(s,a,z_h^{i/N}) - (\mathcal{T}_h^{\pi_t^{i/N}} \hat{Q}_{h+1}^{\lambda,i})(s,a)\right|^2\right]$$

$$\leq C_1^2 \cdot \mathbb{E}_{\rho_{h+1}^{\mathrm{b},i}}\left[\left|Q_{h+1}^{\lambda,i}(s,a) - \hat{Q}_{h+1}^{\lambda,i}(s,a)\right|^2\right], \tag{F.2}$$

where the equality follows from the definition of operator $\mathcal{T}$, and the inequality follows from Assumption 8. In the follow, we control the second term in the right-hand side of inequality (F.1). For any function $g \in \mathcal{F}_{h+1}$, we define the value function at time $h+1$ for $\pi_t^{i/N}$ induced by g as

$$V_g^i(s) = \langle g(s,\cdot), \pi_{h+1}^{i/N}(\cdot\,|\,s)\rangle + \lambda R\left(\pi_{h+1}^{i/N}(\cdot\,|\,s)\right).$$

19

Then the value function defined in Line 8 of Algorithm 2 can be expressed as $\hat{V}_{h+1}^{\lambda,i} = V_{\hat{Q}_{h+1}^{\lambda,i}}^i$. For any function $g \in \mathcal{F}_{h+1}$, we define the regression problem and the corresponding estimate

$$\hat{Q}_g = \operatorname*{argmin}_{f \in \mathcal{F}_h} \sum_{\tau=1}^{K} \left(f(s_{\tau,h}^i, a_{\tau,h}^i) - r_{\tau,h}^i - V_g^i(s_{\tau,h+1}^i) \right)^2.$$

Then $\hat{Q}_h^{\lambda,i} = \hat{Q}_{\hat{Q}_{h+1}^{\lambda,i}}$. Thus, the second term on the right-hand side of inequality F.1 can be bounded as

$$\mathbb{E}_{\rho_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + (\mathcal{P}_h \hat{V}_{h+1}^{\lambda,i})(s, a) - \hat{Q}_h^{\lambda,i}(s, a) \right|^2 \right]$$
$$\le \sup_{g \in \mathcal{F}_{h+1}} \mathbb{E}_{\rho_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + (\mathcal{P}_h V_g^i)(s, a) - \hat{Q}_g(s, a) \right|^2 \right].$$

The term inside the supremum can be handled as

$$\mathbb{E}_{\rho_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + (\mathcal{P}_h V_g^i)(s, a) - \hat{Q}_g(s, a) \right|^2 \right]$$
$$= \mathbb{E}_{\tilde{\rho}_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + V_g^i(s') - \hat{Q}_g(s, a) \right|^2 \right]$$
$$- \mathbb{E}_{\tilde{\rho}_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + (\mathcal{P}_h V_g^i)(s, a) - r_h(s, a, z_h^{i/N}) - V_g^i(s') \right|^2 \right],$$

which follows from the basic calculation. For any $g \in \mathcal{F}_{h+1}$, Assumption 6 implies that

$$r_h(\cdot, \cdot, z_h^{i/N}) + (\mathcal{P}_h V_g^i) \in \mathcal{F}_h.$$

Thus, the definition of $\hat{Q}_g$ shows that

$$\sum_{\tau=1}^{K} \left(r_{\tau,h}^i + V_g^i(s_{\tau,h+1}^i) - \hat{Q}_g(s_{\tau,h}^i, a_{\tau,h}^i) \right)^2 \le \sum_{\tau=1}^{K} \left(r_{\tau,h}^i + V_g^i(s_{\tau,h+1}^i) - r_{\tau,h}^i - (\mathcal{P}_h V_g^i)(s_{\tau,h}^i, a_{\tau,h}^i) \right)^2.$$

Further, we have that

$$\sup_{g \in \mathcal{F}_{h+1}} \mathbb{E}_{\rho_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + (\mathcal{P}_h V_g^i)(s, a) - \hat{Q}_g(s, a) \right|^2 \right]$$
$$\le \sup_{g \in \mathcal{F}_{h+1}} \left\{ \mathbb{E}_{\tilde{\rho}_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + V_g^i(s') - \hat{Q}_g(s, a) \right|^2 \right] - \mathbb{E}_{\tilde{\rho}_h^{b,i}} \left[\left| (\mathcal{P}_h V_g^i)(s, a) - V_g^i(s') \right|^2 \right] \right.$$
$$\left. - \frac{2}{K} \left[\sum_{\tau=1}^{K} \left(r_{\tau,h}^i + V_g^i(s_{\tau,h+1}^i) - \hat{Q}_g(s_{\tau,h}^i, a_{\tau,h}^i) \right)^2 - \sum_{\tau=1}^{K} \left(V_g^i(s_{\tau,h+1}^i) - (\mathcal{P}_h V_g^i)(s_{\tau,h}^i, a_{\tau,h}^i) \right)^2 \right] \right\}$$
$$\le \sup_{g \in \mathcal{F}_{h+1}, f \in \mathcal{F}_h} \left\{ \mathbb{E}_{\tilde{\rho}_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + V_g^i(s') - f(s, a) \right|^2 \right] - \mathbb{E}_{\tilde{\rho}_h^{b,i}} \left[\left| (\mathcal{P}_h V_g^i)(s, a) - V_g^i(s') \right|^2 \right] \right.$$
$$\left. - \frac{2}{K} \left[\sum_{\tau=1}^{K} \left(r_{\tau,h}^i + V_g^i(s_{\tau,h+1}^i) - f(s_{\tau,h}^i, a_{\tau,h}^i) \right)^2 - \sum_{\tau=1}^{K} \left(V_g^i(s_{\tau,h+1}^i) - (\mathcal{P}_h V_g^i)(s_{\tau,h}^i, a_{\tau,h}^i) \right)^2 \right] \right\}.$$

We define that

$$e_{g,f}(s, a, s') = \left(r_h(s, a, z_h^{i/N}) + V_g^i(s') - f(s, a) \right)^2 - \left(V_g^i(s') - (\mathcal{P}_h V_g^i)(s, a) \right)^2.$$

Then we have that

$$\mathbb{E}_{\rho_h^{b,i}} \left[\left| r_h(s, a, z_h^{i/N}) + (\mathcal{P}_h \hat{V}_{h+1}^{\lambda,i})(s, a) - \hat{Q}_h^{\lambda,i}(s, a) \right|^2 \right]$$
$$\le \sup_{g \in \mathcal{F}_{h+1}, f \in \mathcal{F}_h} \left\{ \mathbb{E}_{\tilde{\rho}_h^{b,i}} \left[e_{g,f}(s, a, s') \right] - \frac{2}{K} \sum_{\tau=1}^{K} e_{g,f}(s_{\tau,h}^i, a_{\tau,h}^i, s_{\tau,h+1}^i) \right\}.$$

Define $B_H = (1 + \lambda \log |\mathcal{A}|) H$. The bound for the generalization error is as follows.

Proposition 7. *For any $\varepsilon, \gamma, \theta > 0$ we have*

$$\mathbb{P}\bigg(\exists f \in \mathcal{F}_h, g \in \mathcal{F}_{h+1}, \mathbb{E}_{\tilde{\rho}_h^{b,i}} \big[e_{g,f}(s,a,s') \big] - \frac{1}{K} \sum_{\tau=1}^{K} e_{g,f}(s_{\tau,h}^i, a_{\tau,h}^i, s_{\tau,h+1}^i) $$

$$\geq \varepsilon \Big(\gamma + \theta + \mathbb{E}_{\tilde{\rho}_h^{b,i}} \big[e_{g,f}(s,a,s') \big] \Big) \bigg)$$

$$\leq 12 \mathcal{N}_\infty \bigg(\frac{\varepsilon \theta}{320 B_H^3}, \mathcal{F}_h \bigg) \cdot \mathcal{N}_\infty \bigg(\frac{\varepsilon \theta}{320 B_H^3}, \mathcal{F}_{h+1} \bigg) \cdot \exp\bigg(- \frac{\varepsilon^2 (1-\varepsilon) \gamma K}{280(1+\varepsilon) B_H^4} \bigg).$$

Proof of Proposition 7. See Appendix I.5. $\qquad \square$

We take $\varepsilon = 1/2$, $\gamma = \theta = t/2$, then we have

$$\mathbb{P}\bigg(\sup_{g \in \mathcal{F}_{h+1}, f \in \mathcal{F}_h} \Big\{ \mathbb{E}_{\tilde{\rho}_h^{b,i}} \big[e_{g,f}(s,a,s') \big] - \frac{2}{K} \sum_{\tau=1}^{K} e_{g,f}(s_{\tau,h}^i, a_{\tau,h}^i, s_{\tau,h+1}^i) \Big\} \geq t \bigg)$$

$$\leq 12 \mathcal{N}_\infty \bigg(\frac{t}{1280 B_H^3}, \mathcal{F}_h \bigg) \cdot \mathcal{N}_\infty \bigg(\frac{t}{1280 B_H^3}, \mathcal{F}_{h+1} \bigg) \cdot \exp\bigg(- \frac{tK}{6720 B_H^4} \bigg).$$

Thus, with probability at least $1 - \delta$, we have

$$\mathbb{E}_{\rho_h^{b,i}} \Big[\big| r_h(s,a,z_h^{i/N}) + (\mathcal{P}_h \hat{V}_{h+1}^{\lambda,i})(s,a) - \hat{Q}_h^{\lambda,i}(s,a) \big|^2 \Big]$$

$$= O\bigg(\frac{B_H^4}{K} \log \frac{\mathcal{N}_\infty(5 B_H / K, \mathcal{F}_h) \mathcal{N}_\infty(5 B_H / K, \mathcal{F}_{h+1})}{\delta} \bigg)$$

Substituting this inequality and inequality (F.2) into inequality (F.1), we have that

$$\mathbb{E}_{\rho_h^{b,i}} \Big[\big| Q_h^{\lambda,i}(s,a) - \hat{Q}_h^{\lambda,i}(s,a) \big|^2 \Big]$$

$$\leq C_1^2 \cdot \mathbb{E}_{\rho_{h+1}^{b,i}} \Big[\big| Q_{h+1}^{\lambda,i}(s,a) - \hat{Q}_{h+1}^{\lambda,i}(s,a) \big|^2 \Big]$$

$$+ O\bigg(\frac{B_H^4}{K} \log \frac{\mathcal{N}_\infty(5 B_H / K, \mathcal{F}_h) \mathcal{N}_\infty(5 B_H / K, \mathcal{F}_{h+1})}{\delta} \bigg).$$

Define the maximal covering number $\mathcal{N}_\infty(\delta, \mathcal{F}_{[H]}) = \max_{h \in [H]} \mathcal{N}_\infty(\delta, \mathcal{F}_h)$. Then from the union bound, we have that with probability at least $1 - \delta$, for any $i \in [N], h \in [H]$

$$\mathbb{E}_{\rho_h^{b,i}} \Big[\big| Q_h^{\lambda,i}(s,a) - \hat{Q}_h^{\lambda,i}(s,a) \big|^2 \Big] = O\bigg(C_1^2 \frac{H B_H^4}{K} \log \frac{NH \cdot \mathcal{N}_\infty(5 B_H / K, \mathcal{F}_{[H]})}{\delta} \bigg).$$

Step 2: Combine the estimation result with the optimization result in Theorem 3.

From the proof of Theorem 3, we need to bound the term ε_h^α. We divide the interval $\mathcal{I} = [0,1]$ into N small intervals $\mathcal{I}_i = ((i-1)/N, i/N]$ for $i \in \{2, \cdots, N\}$ and $\mathcal{I}_1 = [0, i/N]$. For any $\alpha \in \mathcal{I}_i$, we have that

$$\Big| \langle \hat{Q}_h^{\lambda,\alpha}(s_h^\alpha, \cdot, \pi_t^\alpha, \mu_t^{\mathcal{I}}) - Q_h^{\lambda,\alpha}(s_h^\alpha, \cdot, \pi_t^\alpha, \mu_t^{\mathcal{I}}), \pi_h^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha) \rangle \Big|$$

$$\leq \Big| \langle \hat{Q}_h^{\lambda,i}(s_h^\alpha, \cdot) - Q_h^{\lambda,i}(s_h^\alpha, \cdot), \pi_h^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha) \rangle \Big|$$

$$+ \Big| \langle Q_h^{\lambda,i}(s_h^\alpha, \cdot) - Q_h^{\lambda,\alpha}(s_h^\alpha, \cdot, \pi_t^\alpha, \mu_t^{\mathcal{I}}), \pi_h^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha) \rangle \Big|$$

$$\leq 2 C_1 \Big| \langle \hat{Q}_h^{\lambda,i}(s_h^\alpha, \cdot) - Q_h^{\lambda,i}(s_h^\alpha, \cdot), \pi_{t,h}^{b,i}(\cdot \mid s_h^\alpha) \rangle \Big| + O\Big(\frac{\lambda \log T}{N} \Big),$$

where the first inequality results from the triangle inequality, the second inequality results from Assumption 8, the Lipschitzness of reward function in Assumption **??**, the Lipschiz constant of R for distributions $p \geq \mathrm{Unif}(\mathcal{A})/T$, and Cauchy–Schwarz inequality, and we omit the Lipschitz constant

21

dependency on L_r for ease of notation. Then with probability at least $1 - \delta$, the first term on the right-hand side of this inequality can be controlled as

$$\mathbb{E}_{\mu_h^{*,\alpha}}\left[\left|\langle \hat{Q}_h^{\lambda,i}(s_h^\alpha, \cdot) - Q_h^{\lambda,i}(s_h^\alpha, \cdot), \pi_{t,h}^{\text{b},i}(\cdot \mid s_h^\alpha)\rangle\right|\right]$$

$$\leq C_2 \mathbb{E}_{\rho_h^{\text{b},i}}\left[\left|\langle \hat{Q}_h^{\lambda,i}(s_h^\alpha, \cdot) - Q_h^{\lambda,i}(s_h^\alpha, \cdot), \pi_{t,h}^{\text{b},i}(\cdot \mid s_h^\alpha)\rangle\right|\right]$$

$$\leq C_2 \sqrt{\mathbb{E}_{\rho_h^{\text{b},i}}\left[\left(\hat{Q}_h^{\lambda,i}(s, a) - Q_h^{\lambda,i}(s, a)\right)^2\right]}$$

$$\leq C_1 C_2 \frac{\sqrt{H} B_H^2}{\sqrt{K}} \log \frac{TNH \cdot \mathcal{N}_\infty(5B_H/K, \mathcal{F}_{[H]})}{\delta},$$

where the first inequality results from Assumption 8, the second inequality results from Hölder inequality, and the last inequality results from Step 1 and the union bound for $t \in [T]$. Thus, we have

$$\int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}[\varepsilon_h^\alpha]\mathrm{d}\alpha = O\left(C_1 C_2 \frac{H^{3/2} B_H^2}{\sqrt{K}} \log \frac{TNH \cdot \mathcal{N}_\infty(5B_H/K, \mathcal{F}_{[H]})}{\delta} + \frac{\lambda H \log T}{N}\right).$$

Combined with the proof of Theorem 3, this concludes the proof of Theorem 4. $\qquad \square$

G Proof of Proposition 1

Proof of Proposition 1. The existence of the NE follows from Theorem 1. Here we only prove that there are at most one NE. Suppose there exists two different NEs $(\pi^\mathcal{I}, \mu^\mathcal{I})$ and $(\tilde{\pi}^\mathcal{I}, \tilde{\mu}^\mathcal{I})$. According to the definition of NE, we have that

$$\int_0^1 J^{\lambda,\alpha}(\pi^\alpha, \mu^\mathcal{I}) - J^{\lambda,\alpha}(\tilde{\pi}^\alpha, \mu^\mathcal{I})\mathrm{d}\alpha \geq 0$$

$$\int_0^1 J^{\lambda,\alpha}(\tilde{\pi}^\alpha, \tilde{\mu}^\mathcal{I}) - J^{\lambda,\alpha}(\pi^\alpha, \tilde{\mu}^\mathcal{I})\mathrm{d}\alpha \geq 0.$$

Summing these two inequalities, we have

$$\int_0^1 J^{\lambda,\alpha}(\pi^\alpha, \mu^\mathcal{I}) + J^{\lambda,\alpha}(\tilde{\pi}^\alpha, \tilde{\mu}^\mathcal{I}) - J^{\lambda,\alpha}(\tilde{\pi}^\alpha, \mu^\mathcal{I}) - J^{\lambda,\alpha}(\pi^\alpha, \tilde{\mu}^\mathcal{I})\mathrm{d}\alpha \geq 0,$$

which contradicts the strictly weak monotone condition. $\qquad \square$

H Proof of Proposition 2

Proof of Proposition 2. We first note that

$$J^{\lambda,\alpha}(\pi^\alpha, \mu^\mathcal{I}) - J^{\lambda,\alpha}(\pi^\alpha, \tilde{\mu}^\mathcal{I}) = J^\alpha(\pi^\alpha, \mu^\mathcal{I}) - J^\alpha(\pi^\alpha, \tilde{\mu}^\mathcal{I}),$$

$$J^{\lambda,\alpha}(\tilde{\pi}^\alpha, \tilde{\mu}^\mathcal{I}) - J^{\lambda,\alpha}(\tilde{\pi}^\alpha, \mu^\mathcal{I}) = J^\alpha(\tilde{\pi}^\alpha, \tilde{\mu}^\mathcal{I}) - J^\alpha(\tilde{\pi}^\alpha, \mu^\mathcal{I}),$$

since the transition kernel is independent of the distribution flow, where we denote $J^{\lambda,\alpha}(\pi^\alpha, \mu^\mathcal{I})$ for $\lambda = 0$ as $J^\alpha(\pi^\alpha, \mu^\mathcal{I})$. Thus, the desired inequality is equivalent to

$$\int_0^1 J^\alpha(\pi^\alpha, \mu^\mathcal{I}) + J^\alpha(\tilde{\pi}^\alpha, \tilde{\mu}^\mathcal{I}) - J^\alpha(\tilde{\pi}^\alpha, \mu^\mathcal{I}) - J^\alpha(\pi^\alpha, \tilde{\mu}^\mathcal{I})\mathrm{d}\alpha \leq 0.$$

We define $\rho_h^\mathcal{I}, \tilde{\rho}_h^\mathcal{I} \in \Delta(\mathcal{S} \times \mathcal{A})^\mathcal{I}$ for $h \in [H]$ as $\rho_h^\alpha(s, a) = \mu_h^\alpha(s)\pi_h^\alpha(a \mid s)$ and $\tilde{\rho}_h^\alpha(s, a) = \tilde{\mu}_h^\alpha(s)\tilde{\pi}_h^\alpha(a \mid s)$ for all $h \in [H]$ and $\alpha \in \mathcal{I}$. Then the weakly monotone condition implies that

$$\int_\mathcal{I} \sum_{a \in \mathcal{A}} \int_\mathcal{S} \left(\rho_h^\alpha(s, a) - \tilde{\rho}_h^\alpha(s, a)\right)\left(r_h\big(s, a, z_h^\alpha(\mu_h^\mathcal{I}, W_h)\big) - r_h\big(s, a, z_h^\alpha(\tilde{\mu}_h^\mathcal{I}, W_h)\big)\right)\mathrm{d}s\mathrm{d}\alpha \leq 0.$$

Then we have that

$$\int_0^1 J^\alpha(\pi^\alpha, \mu^\mathcal{I}) + J^\alpha(\tilde{\pi}^\alpha, \tilde{\mu}^\mathcal{I}) - J^\alpha(\tilde{\pi}^\alpha, \mu^\mathcal{I}) - J^\alpha(\pi^\alpha, \tilde{\mu}^\mathcal{I})\mathrm{d}\alpha$$

$$= \sum_{h=1}^{H} \int_{\mathcal{I}} \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \left(\rho_h^\alpha(s,a) - \tilde{\rho}_h^\alpha(s,a) \right) \left(r_h \left(s, a, z_h^\alpha(\mu_h^{\mathcal{I}}, W_h) \right) - r_h \left(s, a, z_h^\alpha(\tilde{\mu}_h^{\mathcal{I}}, W_h) \right) \right) \mathrm{d}s \mathrm{d}\alpha$$

$$\leq 0.$$

Thus, we conclude the proof of Proposition 2. $\qquad\square$

I Supporting Propositions and Lemmas

I.1 Proof of Proposition 3

Proof of Proposition 3. We first prove that $T_h^\nu u \in C_h(\mathcal{S})$ for any $u \in C_{h+1}(\mathcal{S})$. By Proposition 7.32 in Bertsekas and Shreve [1996], $T_h^\nu u$ is continuous. The sup-norm of it can be upper-bounded as

$$\|T_h^\nu u\|_\infty \leq 1 + \lambda \log |\mathcal{A}| + (H - h)(1 + \lambda \log |\mathcal{A}|) = (H - h + 1)(1 + \lambda \log |\mathcal{A}|).$$

For the second claim, we have that

$$\|T_h^\nu u - T_h^\nu v\|_\infty$$

$$= \sup_{s \in \mathcal{S}} \Big| \sup_{p \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) \bar{r}_h(s, a, \bar{\nu}_h) - \lambda R(p) + \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} p(a) \bar{P}_h(s' \mid s, a, \bar{\nu}_h) u(s') \mathrm{d}s'$$

$$- \sup_{q \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} q(a) \bar{r}_h(s, a, \bar{\nu}_h) - \lambda R(q) + \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} q(a) \bar{P}_h(s' \mid s, a, \bar{\nu}_h) v(s') \mathrm{d}s' \Big|$$

$$\leq \sup_{s \in \mathcal{S}, p \in \Delta(\mathcal{A})} \Big| \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} p(a) \bar{P}_h(s' \mid s, a, \bar{\nu}_h) \left(v(s') - u(s') \right) \mathrm{d}s' \Big|$$

$$\leq \|u - v\|_\infty,$$

where the first inequality results from that $|\sup_{x \in \mathcal{X}} f(x) - \sup_{x \in \mathcal{X}} g(x)| \leq \sup_{x \in \mathcal{X}} |f(x) - g(x)|$ for any real-valued functions f, g and set $\mathcal{X}$. Thus, we conclude the proof of Proposition 3. $\qquad\square$

I.2 Proof of Proposition 6

Proof of Proposition 6. Let $\{(\nu^{(n)}, \xi^{(n)})\}_{n \geq 1} \subseteq \Xi \times \Xi$ be a sequence such that $\xi^{(n)} \in \Gamma(\nu^{(n)})$ for all $n \geq 1$ and $(\nu^{(n)}, \xi^{(n)}) \to (\nu, \xi)$ as $n \to \infty$ with respect to the total variation distance for some $(\nu, \xi) \in \Xi \times \Xi$. To prove the graph of Γ is closed, we need to prove that $\xi \in \Gamma(\nu)$.

We first prove that $\xi \in A(\nu)$. For any $n \geq 1$ and $h \in [H]$, we have

$$\bar{\xi}_{h+1}^{(n)}(s') = \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \bar{P}_h(\cdot \mid s, a, \bar{\nu}_h^{(n)}) \nu_h^{(n)}(\mathrm{d}s, a).$$

Since $\xi^{(n)} \to \xi$ in total variation, $\xi_h^{(n)} \to \xi_h$ weakly. Take any bounded continuous function $g \in C_b(\mathcal{S})$. Then

$$\lim_{n \to \infty} \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \int_{\mathcal{S}} g(s') \bar{P}_h(\mathrm{d}s' \mid s, a, \bar{\nu}_h^{(n)}) \nu_h^{(n)}(\mathrm{d}s, a) = \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \int_{\mathcal{S}} g(s') \bar{P}_h(\mathrm{d}s' \mid s, a, \bar{\nu}_h) \nu_h(\mathrm{d}s, a),$$

$$\text{(I.1)}$$

which results from Langen [1981], $\int_{\mathcal{S}} g(s') \bar{P}_h(\mathrm{d}s' \mid s, a, \bar{\nu}_h^{(n)})$ converges to $\int_{\mathcal{S}} g(s') \bar{P}_h(\mathrm{d}s' \mid s, a, \bar{\nu}_h)$ continuously, and $\nu^{(n)}$ converges to ν. Eqn. (I.1) implies that $\sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \bar{P}_h(\cdot \mid s, a, \bar{\nu}_h^{(n)}) \nu_h^{(n)}(\mathrm{d}s, a)$ weakly converges to $\sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \bar{P}_h(\cdot \mid s, a, \bar{\nu}_h) \nu_h(\mathrm{d}s, a)$. Thus, we have

$$\bar{\xi}_{h+1}(\cdot) = \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \bar{P}_h(\cdot \mid s, a, \bar{\nu}_h) \nu_h(\mathrm{d}s, a).$$

We then prove that $\xi \in B(\nu)$. Since $\xi^{(n)} \in B(\nu^{(n)})$, there exists sets $A_h^{(n)} \subseteq \mathcal{S}$ for all $n \geq 1$ and $h \in [H]$ such that $\bar{\xi}_h^{(n)}(A_h^{(n)}) = 1$, and for any $n \geq 1$, $h \in [H]$ and $s \in A_h^{(n)}$, the following equation

holds

$$\sum_{a\in\mathcal{A}}\xi_h^{(n)}(a\,|\,s)\bar{r}_h(s,a,\bar{\nu}_h^{(n)}) - \lambda R\big(\xi_h^{(n)}(\cdot\,|\,s)\big) + \sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\xi_h^{(n)}(a\,|\,s)\bar{P}_h(s'\,|\,s,a,\bar{\nu}_h^{(n)})\bar{V}_{h+1}^{\lambda,\nu^{(n)}}(s')\mathrm{d}s' = \bar{V}_h^{\lambda,\nu^{(n)}}(s).$$

$$(\text{I.2})$$

We construct the set $A_h = \cap_{k=1}^{\infty}\cup_{n=k}^{\infty}A_h^{(n)}$ for all $h\in[H]$. Then the following proposition shows that $\bar{\xi}_h(A_h) = 1$ for all $h\in[H]$.

Proposition 8. *Suppose that $\{P_n\}_{n=1}^{\infty}$ and P are distributions on a measurable space, and $P_n \to P$ with respect to the total variation distance as $n\to\infty$. Take any sequence of sets $\{A_n\}_{n=1}^{\infty}$ such that $P_n(A_n) = 1$. Then we have*

$$P\big(\cap_{k=1}^{\infty}\cup_{n=k}^{\infty}A_n\big) = 1.$$

Proof of Proposition 8. See Appendix I.3. □

We are going to prove that for any $s\in A_h$, the following equation holds,

$$\sum_{a\in\mathcal{A}}\xi_h(a\,|\,s)\bar{r}_h(s,a,\bar{\nu}_h) - \lambda R\big(\xi_h(\cdot\,|\,s)\big) + \sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\xi_h(a\,|\,s)\bar{P}_h(s'\,|\,s,a,\bar{\nu}_h)\bar{V}_{h+1}^{\lambda,\nu}(s')\mathrm{d}s' = \bar{V}_h^{\lambda,\nu}(s).$$

We first show that the optimal value functions $\bar{V}_h^{\lambda,\nu^{(n)}}$ converge to $\bar{V}_h^{\lambda,\nu}$ continuously.

Proposition 9. *Given $\mathcal{S}$ is compact, if $\nu^{(n)} \to \nu$ in total variation, we have*

$$\lim_{n\to\infty}\sup_{s\in\mathcal{S}}\big|\bar{V}_h^{\lambda,\nu^{(n)}}(s) - \bar{V}_h^{\lambda,\nu}(s)\big| = 0 \text{ for all } h\in[H].$$

Proof of Proposition 9. See Appendix I.4. □

From the definition of A_h, for any $s\in A_h$, there is a sequence $\{n_k\}_{k=1}^{\infty}$ such that $s\in A_h^{(n_k)}$ for all $k\geq 1$. Since $\xi_h^{(n_k)} \to \xi_h$ in total variation, $\xi_h^{(n_k)}(a\,|\,s) \to \xi_h(a\,|\,s)$ for all $a\in\mathcal{A}$ and $s\in\mathrm{supp}(\bar{\xi}_h)$. Thus, we have

$$\lim_{k\to\infty}\sum_{a\in\mathcal{A}}\xi_h^{(n)}(a\,|\,s)\bar{r}_h(s,a,\bar{\nu}_h^{(n)}) - \lambda R\big(\xi_h^{(n)}(\cdot\,|\,s)\big) + \sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\xi_h^{(n)}(a\,|\,s)\bar{P}_h(s'\,|\,s,a,\bar{\nu}_h^{(n)})\bar{V}_{h+1}^{\lambda,\nu^{(n)}}(s')\mathrm{d}s'$$

$$= \sum_{a\in\mathcal{A}}\xi_h(a\,|\,s)\bar{r}_h(s,a,\bar{\nu}_h) - \lambda R\big(\xi_h(\cdot\,|\,s)\big) + \sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\xi_h(a\,|\,s)\bar{P}_h(s'\,|\,s,a,\bar{\nu}_h)\bar{V}_{h+1}^{\lambda,\nu}(s')\mathrm{d}s', \quad (\text{I.3})$$

which results from Langen [1981], Assumption 4 (3) and (4), Proposition 9 and $\xi_h^{(n_k)}(a\,|\,s) \to \xi_h(a\,|\,s)$. Combining Eqn. (I.3) and that $\bar{V}_{h+1}^{\lambda,\nu^{(n_k)}}(s) \to \bar{V}_{h+1}^{\lambda,\nu}(s)$, we prove the Eqn. (I.2). Thus, we conclude the proof of Proposition 6. □

I.3 Proof of Proposition 8

Proof of Proposition 8. Define the event $G = \cap_{k=1}^{\infty}\cup_{n=k}^{\infty}A_n$ and $B_k = \cap_{n=k}^{\infty}A_n^{\complement}$. Then we have that $G^{\complement} = \cup_{k=1}^{\infty}B_k$. Note that $B_k \subseteq B_{k+1}$ and the monotone convergence theorem, we have that

$$P_n\big(G^{\complement}\cap A_n\big) = \liminf_{k\to\infty}P_n\big(B_k\cap A_n\big). \quad (\text{I.4})$$

We then have that

$$1 = \limsup_{n\to\infty}\liminf_{k\to\infty}\Big[P_n\big(G\cap A_n\big) + P_n\big(B_k\cap A_n\big)\Big]$$

$$\leq \liminf_{k\to\infty}\limsup_{n\to\infty}\Big[P_n\big(G\cap A_n\big) + P_n\big(B_k\cap A_n\big)\Big], \quad (\text{I.5})$$

where the equation results from Eqn. (I.4). For the second term in the right-hand side of inequality (I.5), we fix any $k > 0$, then for $n > k$, we have that $P_n(B_k \cap A_n) = 0$ from the definition of B_k. Thus, we have that

$$\limsup_{n\to\infty} P_n(B_k \cap A_n) = 0, \liminf_{k\to\infty} \limsup_{n\to\infty} P_n(G \cap A_n) = 1,$$

where the second equation results from the first equation and inequality (I.5). Since P_n converges to P in total variation distance, the weak convergence of P_n to P is guaranteed. Portmanteau Theorem shows that

$$P(G) \geq \limsup_{n\to\infty} P_n(G) \geq \liminf_{k\to\infty} \limsup_{n\to\infty} P_n(G \cap A_n) = 1.$$

This concludes the proof of Proposition 8. $\qquad\square$

I.4 Proof of Proposition 9

Proof of Proposition 9. For ease of notation, we define $T^{(n)} = T^{\nu^{(n)}}$, $T = T^\nu$, $\bar{V}^{\lambda,(n)} = \bar{V}^{\lambda,\nu^{(n)}}$, $\bar{V}^\lambda = \bar{V}^{\lambda,\nu}$, $u_0^{(n)} = u_0 = 0$ and we let

$$u_{k+1}^{(n)} = T^{(n)} u_k^{(n)} \quad u_{k+1} = T u_k \text{ for } k \geq 1.$$

From the contraction property in inequality (D.1), we have

$$\rho(u_k^{(n)}, \bar{V}^{\lambda,(n)}) \leq \sigma^k L_0 \quad \rho(u_k, \bar{V}^\lambda) \leq \sigma^k L_0.$$

We then prove that $\lim_{n\to\infty} \sup_{s\in\mathcal{S}} |u_{k,h}^{(n)}(s) - u_{k,h}(s)| = 0$ for all $h \in [H]$ and $k \geq 0$. We prove this by induction. When $k = 0$, $u_{0,h}^{(n)}(s) = u_{0,h}^{(n)}(s) = 0$ from definition. Suppose that the claim holds for k and all $h \in [H]$. Consider $k + 1$ and any $h \in [H]$, we have that

$$\sup_{s\in\mathcal{S}} |u_{k+1,h}^{(n)}(s) - u_{k+1,h}(s)|$$

$$= \sup_{s\in\mathcal{S}} \Big| \sup_{p\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} p(a)\bar{r}_h(s,a,\bar{\nu}_h^{(n)}) - \lambda R(p) + \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} p(a)\bar{P}_h(s' \mid s,a,\bar{\nu}_h^{(n)}) u_{k,h+1}^{(n)}(s')\mathrm{d}s'$$

$$- \sup_{q\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} q(a)\bar{r}_h(s,a,\bar{\nu}_h) - \lambda R(q) + \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} q(a)\bar{P}_h(s' \mid s,a,\bar{\nu}_h) u_{k,h+1}(s')\mathrm{d}s' \Big|$$

$$\leq \sup_{s\in\mathcal{S},p\in\Delta(\mathcal{A})} \Big| \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} p(a)\bar{P}_h(s' \mid s,a,\bar{\nu}_h^{(n)}) u_{k,h+1}^{(n)}(s')\mathrm{d}s' - \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} q(a)\bar{P}_h(s' \mid s,a,\bar{\nu}_h) u_{k,h+1}(s')\mathrm{d}s' \Big|$$

$$+ \sup_{s\in\mathcal{S},p\in\Delta(\mathcal{A})} \Big| \sum_{a\in\mathcal{A}} p(a)\big(\bar{r}_h(s,a,\bar{\nu}_h^{(n)}) - \bar{r}_h(s,a,\bar{\nu}_h)\big) \Big|.$$

From Assumption 4 (2), $\bar{r}_h(\cdot,\cdot,\bar{\nu}_h^{(n)})$ converges to $\bar{r}_h(\cdot,\cdot,\bar{\nu}_h)$ continuously. Also, since continuous function $u_{k,h+1}^{(n)}$ uniformly converges to $u_{k,h+1}$ on compact sets $\mathcal{S}$ as $n \to \infty$, we have that $u_{k,h+1}^{(n)}$ converges to $u_{k,h+1}$ continuously. By [Langen, 1981, Theorem 3.5] and Assumption 4 (4), we have that $\int_{\mathcal{S}} \bar{P}_h(s' \mid \cdot,\cdot,\bar{\nu}_h^{(n)}) u_{k,h+1}^{(n)}(s')\mathrm{d}s'$ continuously converges to $\int_{\mathcal{S}} \bar{P}_h(s' \mid \cdot,\cdot,\bar{\nu}_h) u_{k,h+1}(s')\mathrm{d}s'$. Since the continuous convergence is equivalent to uniform convergence on compact sets, we have that $\lim_{n\to\infty} \sup_{s\in\mathcal{S}} |u_{k+1,h}^{(n)}(s) - u_{k+1,h}(s)| = 0$.

Thus, we have that

$$\sup_{s\in\mathcal{S}} \big| \bar{V}_h^{\lambda,\nu^{(n)}}(s) - \bar{V}_h^{\lambda,\nu}(s) \big|$$

$$\leq \|\bar{V}_h^{\lambda,(n)} - u_{k,h}^{(n)}\|_\infty + \|u_{k,h}^{(n)} - u_{k,h}\|_\infty + \|\bar{V}_h^\lambda - u_{k,h}\|_\infty$$

$$\leq \sigma^h \rho(\bar{V}^{\lambda,(n)}, u_k^{(n)}) + \|u_{k,h}^{(n)} - u_{k,h}\|_\infty + \sigma^h \rho(\bar{V}^\lambda, u_k)$$

$$\leq 2\sigma^{h+k} L_0 + \|u_{k,h}^{(n)} - u_{k,h}\|_\infty. \tag{I.6}$$

The right-hand side of inequality (I.6) can be made arbitrarily small by first choosing a large enough k and then let $n \to \infty$. Thus, we conclude the proof of Proposition 9. $\qquad\square$

I.5 Proof of Proposition 7

Proof of Proposition 7. The proof of Proposition 7 generally follows the proof of Theorem 11.4 in Györfi et al. [2002]. Here we only specify the different parts. In the following, we bound the ℓ_1 covering number of function class $\{e_{g,f} \mid f \in \mathcal{F}_h, g \in \mathcal{F}_{h+1}\}$ on samples $\{(s^i_{\tau,h}, a^i_{\tau,h}, s^i_{\tau,h+1})\}^K_{\tau=1}$. Assume that we have the δ-covers $\mathcal{C}^\delta_h$ and $\mathcal{C}^\delta_{h+1}$ of $\mathcal{F}_h$ and $\mathcal{F}_{h+1}$ with respect to the ℓ_∞, i.e., for any $f \in \mathcal{F}_h$, there exists $f_c \in \mathcal{C}^\delta_h$ such that $\|f - f_c\|_\infty = \sup_{s \in \mathcal{S}, a \in \mathcal{A}} |f(s,a) - f_c(s,a)| \leq \delta$. Then for any $e_{g,f}$, we can find $f_c \in \mathcal{C}^\delta_h$ and $g_c \in \mathcal{C}^\delta_{h+1}$ such that $\|f - f_c\|_\infty, \|g - g_c\|_\infty \leq \delta$. The ℓ_1 distance between $e_{g,f}$ and e_{g_c,f_c} on samples $\{(s^i_{\tau,h}, a^i_{\tau,h}, s^i_{\tau,h+1})\}^K_{\tau=1}$ can be bounded as

$$\frac{1}{K}\sum_{\tau=1}^{K}\left|e_{g,f}(s^i_{\tau,h}, a^i_{\tau,h}, s^i_{\tau,h+1}) - e_{g_c,f_c}(s^i_{\tau,h}, a^i_{\tau,h}, s^i_{\tau,h+1})\right|$$

$$\leq \frac{1}{K}\sum_{\tau=1}^{K}\left\{\left|\left(r_h(s^i_{\tau,h}, a^i_{\tau,h}, z^{i/N}_h) + V^i_g(s^i_{\tau,h+1}) - f(s^i_{\tau,h}, a^i_{\tau,h})\right)^2\right.\right.$$

$$\left.- \left(r_h(s^i_{\tau,h}, a^i_{\tau,h}, z^{i/N}_h) + V_{g_c}(s^i_{\tau,h+1}) - f_c(s^i_{\tau,h}, a^i_{\tau,h})\right)^2\right|$$

$$\left.+ \left|\left(V^i_g(s^i_{\tau,h+1}) - (\mathcal{P}_h V^i_g)(s^i_{\tau,h}, a^i_{\tau,h})\right)^2 - \left(V_{g_c}(s^i_{\tau,h+1}) - (\mathcal{P}_h V_{g_c})(s^i_{\tau,h}, a^i_{\tau,h})\right)^2\right|\right\}. \tag{I.7}$$

For the first term in the right-hand side of this inequality can be bounded as

$$\frac{1}{K}\sum_{\tau=1}^{K}\left|\left(r_h(s^i_{\tau,h}, a^i_{\tau,h}, z^{i/N}_h) + V^i_g(s^i_{\tau,h+1}) - f(s^i_{\tau,h}, a^i_{\tau,h})\right)^2\right.$$

$$\left.- \left(r_h(s^i_{\tau,h}, a^i_{\tau,h}, z^{i/N}_h) + V_{g_c}(s^i_{\tau,h+1}) - f_c(s^i_{\tau,h}, a^i_{\tau,h})\right)^2\right|$$

$$\leq \frac{2B_H}{K}\sum_{\tau=1}^{K}\left|V^i_g(s^i_{\tau,h+1}) - V_{g_c}(s^i_{\tau,h+1})\right| + \left|f(s^i_{\tau,h}, a^i_{\tau,h}) - f_c(s^i_{\tau,h}, a^i_{\tau,h})\right|$$

$$\leq 4B_H\delta,$$

where the last inequality results from the definition of f_c and g_c. The second term in the right-hand side of inequality I.7 can be similarly bounded, then we have that

$$\frac{1}{K}\sum_{\tau=1}^{K}\left|e_{g,f}(s^i_{\tau,h}, a^i_{\tau,h}, s^i_{\tau,h+1}) - e_{g_c,f_c}(s^i_{\tau,h}, a^i_{\tau,h}, s^i_{\tau,h+1})\right| \leq 8B_H\delta.$$

The covering number can be correspondingly bounded as

$$\mathcal{N}_1\left(\delta, \{e_{g,f}\}, \{(s^i_{\tau,h}, a^i_{\tau,h}, s^i_{\tau,h+1})\}^K_{\tau=1}\right) \leq \mathcal{N}_\infty(\delta/8B_N, \mathcal{F}_h)\cdot\mathcal{N}_\infty(\delta/8B_N, \mathcal{F}_{h+1}).$$

Combined with the proof of Theorem 11.4 in Györfi et al. [2002], this concludes the proof of Proposition 7. □

Lemma 10. *For any two distributions $p^*, p \in \Delta(\mathcal{A})$ and $\hat{p} = (1-\beta)p + \beta \text{Unif}(\mathcal{A})$ with $\beta \in (0,1)$. Then*

$$D_{\mathrm{kl}}(p^*\|\hat{p}) \leq \log\frac{|\mathcal{A}|}{\beta}$$

$$D_{\mathrm{kl}}(p^*\|\hat{p}) - D_{\mathrm{kl}}(p^*\|p) \leq \beta/(1-\beta).$$

Proof of Lemma 10.

$$D_{\mathrm{kl}}(p^*\|\hat{p}) \leq \left\langle p^*, \log\frac{p^*}{(1-\beta)p + \beta/|\mathcal{A}|}\right\rangle \leq \left\langle p^*, \log\frac{1}{\beta/|\mathcal{A}|}\right\rangle = \log\frac{|\mathcal{A}|}{\beta}.$$

Thus, we prove the first inequality. For the second inequality, we have

$$D_{\mathrm{kl}}(p^*\|\hat{p}) - D_{\mathrm{kl}}(p^*\|p) = \left\langle p^*, \log\frac{p}{(1-\beta)p + \beta/|\mathcal{A}|}\right\rangle \leq \left\langle p^*, \log\frac{p}{(1-\beta)p}\right\rangle \leq \left\langle p^*, \frac{\beta}{1-\beta}\right\rangle = \frac{\beta}{1-\beta},$$

where the second inequality results from $\log(x) \le x - 1$ for $x > 0$. Thus, we conclude the proof of Lemma 10.

$\square$

Lemma 11 (Performance Difference Lemma). *Given a policy $\pi^{\mathcal{I}}$ and the corresponding mean-field flow $\mu^{\mathcal{I}}$, for any player $\alpha \in \mathcal{I}$ and any policy $\tilde{\pi}^{\alpha}$, we have*

$$
V_1^{\lambda,\alpha}(s, \tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}) - V_1^{\lambda,\alpha}(s, \pi^{\alpha}, \mu^{\mathcal{I}}) + \lambda \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} D_{\mathrm{kl}}\big(\tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha}) \| \pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\big) \mid s_1^{\alpha} = s \right]
$$

$$
= \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} \big\langle Q_h^{\lambda,\alpha}(s_h^{\alpha}, \cdot, \pi^{\alpha}, \mu^{\mathcal{I}}) - \lambda \log \pi_h^{\alpha}(\cdot \mid s_h^{\alpha}), \tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha}) - \pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\big\rangle \mid s_1^{\alpha} = s \right],
$$

where the expectation $\mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}$ is taken with respect to the randomness in implementing policy $\tilde{\pi}^{\alpha}$ for player α under the MDP induced by $\mu^{\mathcal{I}}$.

Proof of Lemma 11. From the definition of $V_1^{\lambda,\alpha}(s, \tilde{\pi}^{\alpha}, \mu^{\mathcal{I}})$, we have

$$
V_1^{\lambda,\alpha}(s, \tilde{\pi}^{\alpha}, \mu^{\mathcal{I}})
$$

$$
= \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} r_h(s_h^{\alpha}, a_h^{\alpha}, z_h^{\alpha}) - \lambda \log \tilde{\pi}_h^{\alpha}(a_h^{\alpha} \mid s_h^{\alpha}) + V_h^{\lambda,\alpha}(s_h^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) - V_h^{\lambda,\alpha}(s_h^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) \,\middle|\, s_1^{\alpha} = s \right]
$$

$$
= \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} r_h(s_h^{\alpha}, a_h^{\alpha}, z_h^{\alpha}) - \lambda \log \tilde{\pi}_h^{\alpha}(a_h^{\alpha} \mid s_h^{\alpha}) + V_{h+1}^{\lambda,\alpha}(s_{h+1}^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) \right.
$$

$$
\left. - V_h^{\lambda,\alpha}(s_h^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) \,\middle|\, s_1^{\alpha} = s \right] + V_1^{\lambda,\alpha}(s, \pi^{\alpha}, \mu^{\mathcal{I}}), \tag{I.8}
$$

where the second equation results from the rearrangement from the terms. We then focus on a part of the right-hand side of Eqn. (I.8).

$$
\mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[r_h(s_h^{\alpha}, a_h^{\alpha}, z_h^{\alpha}) - \lambda \log \tilde{\pi}_h^{\alpha}(a_h^{\alpha} \mid s_h^{\alpha}) + V_{h+1}^{\lambda,\alpha}(s_{h+1}^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) \mid s_1^{\alpha} = s \right]
$$

$$
= \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[r_h(s_h^{\alpha}, a_h^{\alpha}, z_h^{\alpha}) + V_{h+1}^{\lambda,\alpha}(s_{h+1}^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) \mid s_1^{\alpha} = s \right] - \lambda \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[R\big(\tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha})\big) \mid s_1^{\alpha} = s \right]
$$

$$
= \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\big\langle Q_h^{\lambda,\alpha}(s_h^{\alpha}, \cdot, \pi^{\alpha}, \mu^{\mathcal{I}}), \tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha})\big\rangle \mid s_1^{\alpha} = s \right] - \lambda \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[R\big(\tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha})\big) \mid s_1^{\alpha} = s \right], \tag{I.9}
$$

where $R(\cdot)$ is the negative entropy function, the inner product $\langle \cdot, \cdot \rangle$ is taken with respect to the action space $\mathcal{A}$, and the second equation results from the definition of $Q_h^{\lambda,\alpha}$ and $V_{h+1}^{\lambda,\alpha}$. Substituting Eqn. (I.9) into Eqn. (I.8) and noting the fact that $V_h^{\lambda,\alpha}(s_h^{\alpha}, \pi^{\alpha}, \mu^{\mathcal{I}}) = \langle Q_h^{\lambda,\alpha}(s_h^{\alpha}, \cdot, \pi^{\alpha}, \mu^{\mathcal{I}}), \pi_h^{\alpha}(\cdot \mid s_h^{\alpha}) \rangle - R(\pi_h^{\alpha}(\cdot \mid s_h^{\alpha}))$, we derive that

$$
V_1^{\lambda,\alpha}(s, \tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}) - V_1^{\lambda,\alpha}(s, \pi^{\alpha}, \mu^{\mathcal{I}})
$$

$$
= \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} \big\langle Q_h^{\lambda,\alpha}(s_h^{\alpha}, \cdot, \pi^{\alpha}, \mu^{\mathcal{I}}), \tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha}) - \pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\big\rangle \mid s_1^{\alpha} = s \right]
$$

$$
- \lambda \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} R\big(\tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha})\big) - R\big(\pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\big) \mid s_1^{\alpha} = s \right]
$$

$$
= \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} \big\langle Q_h^{\lambda,\alpha}(s_h^{\alpha}, \cdot, \pi^{\alpha}, \mu^{\mathcal{I}}), \tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha}) - \pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\big\rangle \mid s_1^{\alpha} = s \right]
$$

$$
- \lambda \mathbb{E}_{\tilde{\pi}^{\alpha}, \mu^{\mathcal{I}}}\left[\sum_{h=1}^{H} D_{\mathrm{kl}}\big(\tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha}) \| \pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\big) + \big\langle \log \pi_h^{\alpha}(\cdot \mid s_h^{\alpha}), \tilde{\pi}_h^{\alpha}(\cdot \mid s_h^{\alpha}) - \pi_h^{\alpha}(\cdot \mid s_h^{\alpha})\big\rangle \mid s_1^{\alpha} = s \right],
$$

where the last equation results from the definition of the negative entropy $R(\cdot)$. This concludes the proof of Lemma 11.

$\square$

Lemma 12 (Lemma 3.3 in Cai et al. [2020]). *For any distribution $p, p^* \in \Delta(\mathcal{A})$ and any function $g : \mathcal{A} \to [0, H]$, it holds for $q \in \Delta(\mathcal{A})$ with $q(\cdot) \propto p(\cdot) \exp\big(\alpha g(\cdot)\big)$ that*

$$\langle g(\cdot), p^*(\cdot) - p(\cdot) \rangle \leq \alpha H^2/2 + \alpha^{-1}\big[D_{\mathrm{kl}}(p^*\|p) - D_{\mathrm{kl}}(p^*\|q)\big].$$