Journal of Machine Learning Research 25 (2024) 1-95 Submitted 10/23; Revised 5/24; Published 11/24

Learning Regularized Graphon Mean-Field Games with Unknown Graphons

Fengzhuo Zhang fzzhang@u.nus.edu
Department of Electrical and Computer Engineering
National University of Singapore
Singapore 117583

Vincent Y. F. Tan vtan@nus.edu.sg
Department of Mathematics
Department of Electrical and Computer Engineering
National University of Singapore
Singapore 119076

Zhaoran Wang zhaoranwang@gmail.com
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208-3109, USA

Zhuoran Yang zhuoranyang.work@gmail.com
Department of Statistics and Data Science
Yale University
New Haven, CT 06511, USA

Editor: Alexandre Proutiere

Abstract

We design and analyze reinforcement learning algorithms for Graphon Mean-Field Games (GMFGs). In contrast to previous works that require the precise values of the graphons, we aim to learn the Nash Equilibrium (NE) of the regularized GMFGs when the graphons are unknown. Our contributions are threefold. First, we propose the Proximal Policy Optimization for GMFG (GMFG-PPO) algorithm and show that it converges at a rate of $\tilde{O}(T^{-1/3})$ after T iterations with an estimation oracle, improving on a previous work by Xie *et al.* (ICML, 2021). Second, using kernel embedding of distributions, we design efficient algorithms to estimate the transition kernels, reward functions, and graphons from sampled agents. Convergence rates are then derived when the positions of the agents are either known or unknown. Results for the combination of the optimization algorithm GMFG-PPO and the estimation algorithm are then provided. These algorithms are the first specifically designed for learning graphons from sampled agents. Finally, the efficacy of the proposed algorithms are corroborated through simulations. These simulations demonstrate that learning the unknown graphons reduces the *exploitability* effectively.

Keywords: Graphon, Mean-Field Games, Multi-Agent Reinforcement Learning, Policy Gradient, Model-Based Reinforcement Learning

1. Introduction

Multi-Agent Reinforcement Learning (MARL) aims to solve sequential decision-making problems in multi-agent systems (Zhang et al., 2021; Gronauer and Diepold, 2022; Oroojlooy and Hajinezhad, 2022). Although MARL has enjoyed tremendous successes across a wide range of real-world applications (Tang and Ha, 2021; Wang et al., 2022a,b; Xu et al., 2021), it suffers from the "curse of many agents" where the sizes of the state and action spaces increase exponentially with the number of agents (Menda et al., 2018; Wang et al., 2020). A potential remedy is to use the *mean-field approximation* (Yang et al., 2018; Carmona et al., 2019). It assumes that the agents are *homogeneous*, and each agent is influenced only by the *common* state distribution of agents. This assumption mitigates the exponential growth of the state and action spaces (Wang et al., 2020; Guo et al., 2022a). However, the homogeneity assumption heavily restricts the applicability of the Mean-Field Game (MFG). For example, analyzing the propagation of Covid-19 in an extremely large population requires modeling the fact that people in different regions have distinct activity intensities. This cannot be captured by the mean-field approximation, which assumes a simplistic *homogeneous* setup. As a result, the Graphon Mean-Field Game (GMFG) is proposed as a means to relax the homogeneity assumption. It captures the heterogeneity of agents through *graphons* and allows the number of agents to be potentially uncountably infinite (Parise and Ozdaglar, 2019; Carmona et al., 2022). GMFGs have achieved great successes in a wide range of applications, including in networks (Gao and Caines, 2019) and epidemics (Aurell et al., 2022a). In Aurell et al. (2022a), the states of people indicate their infection situation, and the graphons represent the propagation intensity between different types of people.

However, learning algorithms for GMFG require significantly more efforts to design and analyze. Cui and Koeppl (2021b) proposed to learn the Nash Equilibrium (NE) of GMFGs by modifying existing MFG learning algorithms. However, these model-free algorithms suffer from the fact that the distribution flow estimation in GMFG requires a large number of samples due to the heterogeneity of the agents. In addition, these algorithms potentially necessitate the use of a very large class of value functions. In particular, this function class should include the nominal value function in GMFG with any graphons to satisfy the realizability assumption (Jin et al., 2021; Zhan et al., 2022). Moreover, existing works only prove the consistency of learning algorithms with rather stringent assumptions (Cui and Koeppl, 2021b; Fabian et al., 2022). These assumptions include the contractivity of the estimated operators and the access to the nominal value functions. The *convergence rates* of algorithms in GMFGs with milder assumptions are currently lacking in the literature.

In this paper, we focus on learning the NE from data collected from a set of sampled agents in a centralized manner. Concretely, the central planner has access to a simulator of the GMFG which generates the states and rewards of agents with the policies of the agent as its inputs. However, only the states and rewards of only a *finite* set of agents are revealed to the learner. Compared with the settings in Cui and Koeppl (2021b) and Fabian et al. (2022), our setting is more relevant in real-world applications where the number of agents is always finite. We aim to learn the NE of the GMFG from the states and rewards of these sampled agents.

Learning the NEs in our problem involves overcoming difficulties from the *statistical* and *optimization* perspectives. From the statistical side, we suffer from the lack of information about the inputs of the functions to estimate. The transition kernels and the reward functions of each agent take as inputs the collective behavior of all the other agents and the graphon. In contrast, we do not know the graphons and only have information provided by a *finite* subset of agents. From the optimization perspective, each agent is faced with a non-stationary environment formed by other agents. Thus, we should design policy optimization procedures that ensure that the policy of each agent converges to the optimal one in a time-varying environment, while also ensuring that the non-stationary environment converges to a NE.

Main Contributions Addressing these difficulties, we summarize our main contributions and results in Table 1 and in more details as follows:

1. We propose and analyze the Proximal Policy Optimization for GMFG (GMFG-PPO) algorithm to learn the NE. Given an estimate oracle, our algorithm implements a Proximal Policy Optimization (PPO)-like algorithm to update the agents' policies (Schulman et al., 2017). The environment is simultaneously updated with a carefully designed learning rate. These strategies overcome the *optimization-related* hurdles. GMFG-PPO achieves a convergence rate $\tilde{O}(T^{-1/3})$, where T is the number of iterations. This convergence rate is faster than that of the algorithm in Xie et al. (2021) and is proved under fewer assumptions. This improvement is attributed to our carefully designed policy and environment update rates. In addition, the analysis of our optimization leads to a faster convergence of the mirror descent algorithm on a fixed MDP. As a byproduct, we generalize the result in Lan (2022) to inhomogeneous MDPs with a finite horizon.

2. We design and analyze the model learning algorithm of GMFG under three different agent sampling schemes, as shown in Table 1. The algorithm first incorporates the graphon with the empirical measure to estimate the mean-embedding of each agent's influence. Then we take this estimate as the input and then perform a regression task; this resolves the *statistical* difficulties mentioned above. In the case where sampled agents have known and fixed positions, Theorem 5 shows that the convergence rate for the model estimate is $O((NL)^{-1} + N^{-1/2})$, where N is the number of sampled agents, and L is the number of samples from each agent. We also consider two additional scenarios—the case in which the agents are *randomly sampled* from the unit interval but their positions are known, and learning from sample agents with *unknown* grid positions. Pertaining to the final scenario, Theorem 7 indicates that the lack of information of the position of the agents results in the sample complexity being degraded by an additional factor of $O(N \log N)$.

3. Our model estimation learning algorithm is the first one proposed for GMFGs. It recovers the underlying graphons from the states sampled from a finite number of agents. This model-learning problem is a considerable generalization of the distribution regression problem (Szabó et al., 2016). Detailed discussions are provided in Section 5.4. Also, our graphon learning setting can be regarded as a novel addition to the existing graphon estimation literature, as discussed in Section 2.

Paper Outline The rest of the paper is organized as follows. We discuss related works in Section 2. In Section 3, we introduce the GMFGs and a key property that they possess,

Table 1: Summary of the theoretical results

Results	Description
Theorem 4	Convergence rate of GMFG-PPO, when an estimation oracle is assumed.
Theorem 5	Convergence rate of the model estimation procedure, when the agents have *known fixed* positions.
Theorem 6	Convergence rate of the model estimation procedure, when the agents have *known random* positions.
Theorem 7	Convergence rate of the model estimation procedure, when the agents have *unknown fixed* positions.
Corollary 9	Convergence rate of the NE learning algorithm that implements GMFG-PPO and collects data from agents with *known fixed* positions.
Corollary 10	Convergence rate of the NE learning algorithm that implements GMFG-PPO and collects data from agents with *known random* positions.
Corollary 11	Convergence rate of the NE learning algorithm that implements GMFG-PPO and collects data from agents with *unknown fixed* positions.

namely *equivariance*. Our three sampling schemes are also introduced. In Section 4, we propose GMFG-PPO and analyze its convergence rate assuming an estimation oracle. In Section 5, we first introduce our mean-embedding procedure. Then we propose and analyze the model-learning algorithms for three sampling schemes. In Section 6, we combine the results from Sections 4 and 5. In Section 7, we provide the numerical simulation results to corroborate our theoretical findings. In Section 8, we conclude our paper.

2. Related Works

The GMFG has been proposed to study the games played between a large number of heterogenous agents for several years. Parise and Ozdaglar (2019) first formulated the static GMFG and proved that it is the limit of finite-agent games with graph structure. Carmona et al. (2022) then generalized these results to the Bayesian setting. Caines and Huang (2019, 2021) formulated the continuous-time GMFG and studied the existence and the uniqueness of their NE. As a special case, the continuous-time linear-quadratic GMFG was studied by Aurell et al. (2022b); Tchuendom et al. (2020); Gao et al. (2020, 2021), where the existence and uniqueness of NE were established, and the convergence of finite-agent games to GMFG was analyzed. Learning of the NE on the discrete-time GMFG was first considered in Vasal et al. (2020) via the master equation. After that, Cui and Koeppl (2021b) and Fabian et al. (2022) proposed algorithms to learn the NE of discrete-time GMFG with dense and sparse graphons, respectively.

As a special case, the MFG models a game between a large number of homogeneous agents. This classical problem formulation was suggested in Lasry and Lions (2007); Huang et al. (2006). NE learning algorithms for the continuous-time MFG have been designed via fictitious play (Cardaliaguet and Hadikhanloo, 2017), mirror descent (Hadikhanloo, 2017), generalized conditional gradient (Lavigne and Pfeiffer, 2022), and policy gradient (Guo et al., 2022b). For discrete-time MFG, efficient algorithms have been proposed based on the notion of contraction (Guo et al., 2019; Xie et al., 2021; Anahtarci et al., 2022; Yardim et al., 2022; Guo et al., 2023). With the monotonicity condition, Perrin et al. (2020) and Perolat et al. (2021) propose fictitious play and mirror descent algorithms for learning the NE, respectively. Readers are encouraged to refer to Laurière et al. (2022) for a comprehensive survey of MFGs.

The graphon estimation problem has been studied for a decade under different classes of graphons and different performance metrics. Existing works mainly focus on the estimation of graphons from the random graphs generated from it. Gao et al. (2015) first proposed a rate-optimal algorithm to estimate the graphon at sampled points. The graphon estimation is then studied under L_2 norm (Klopp et al., 2017; Wolfe and Olhede, 2013), and cut distance (Klopp and Verzelen, 2019). The spectral method for graphon estimation was also studied in Xu (2018). For a comprehensice survey of graphon estimation, readers are encouraged to refer to Gao and Ma (2021). Different from these works, we aim to estimate the graphons without the graphs generated from them. Instead, we only have access to the state and action samples of agents, who interact with each other according to an unknown graphon structure.

Notations We denote $\{1, \cdots, N\}$ as $[N]$. For a set $\mathcal{S}$, we denote the collection of all the measures and the probability measures on $\mathcal{S}$ as $\mathcal{M}(\mathcal{S})$ and $\Delta(\mathcal{S})$, respectively. For a measurable space $(\mathcal{X}, \mathcal{F})$ and two distributions $P, Q \in \Delta(\mathcal{X})$ supported on $\mathcal{X}$, the total variation distance between them is defined as $\mathrm{TV}(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$. For to random variables X, Y supported on $(\mathcal{X}, \mathcal{F})$, we write $\mathrm{TV}(X, Y)$ to denote the total variation between their distributions. For a graphon W, we define its infinity norm as $\|W\|_\infty = \sup_{x,y \in [0,1]} |W(x, y)|$.

3. Preliminaries

Graphons are measurable and *symmetric* functions that map $[0, 1]^2$ to $[0, 1]$. By symmetry, we mean that $W(\alpha, \beta) = W(\beta, \alpha)$ for any $\alpha, \beta \in [0, 1]$. The set of all graphons is denoted as $\mathcal{W} = \{W : [0, 1]^2 \to [0, 1] \,|\, W \text{ is symmetric}\}$. In the following, graphons are used to represent interactions between agents. We consider a finite horizon GMFG $(\mathcal{I}, \mathcal{S}, \mathcal{A}, \mu_1, H, P^*, r^*, W^*)$. In this game, each agent is indexed by $\alpha \in \mathcal{I} = [0, 1]$. The state space and the action space of each agent are respectively denoted as $\mathcal{S} \subseteq \mathbb{R}^{d_{\mathrm{s}}}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_{\mathrm{a}}}$. We assume that $\mathcal{S}$ is a compact subset of $\mathbb{R}^{d_{\mathrm{s}}}$ and $\mathcal{A}$ is a finite subset of $\mathbb{R}^{d_{\mathrm{a}}}$. The horizon of the game is denoted as $H \in \mathbb{N}$. The initial state distribution of each agent is $\mu_1 \in \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the set of probability measures on $\mathcal{S}$. We note that the initial distributions of agents can be different by adding a new time step $h = 0$, where the reward is zero and the transition kernel is independent of the action. The state transition kernels $P^* = \{P_h^*\}_{h=1}^H$ are functions $P_h^* : \mathcal{S} \times \mathcal{A} \times \mathcal{M}(\mathcal{S}) \to \Delta(\mathcal{S})$ for all $h \in [H]$, where $\mathcal{M}(\mathcal{S})$ is the set of measures on $\mathcal{S}$. In contrast to the single-agent Markov Decision Process (MDP), the state dynamics of each agent in a GMFG depends on an *aggregate* $z \in \mathcal{M}(\mathcal{S})$, which reflects the influence of other agents on it. Since we consider the case in which the state space $\mathcal{S}$ is compact but potentially infinite, we assume that $P_h^*(\cdot \,|\, s_h, a_h, z_h)$ admits a probability density function with respect to Lebesgue measure on $\mathcal{S}$ for any $s_h \in \mathcal{S}, a_h \in \mathcal{A}$, $z_h \in \mathcal{M}(\mathcal{S})$, and $h \in [H]$. For time h and agent $\alpha \in \mathcal{I}$, given a graphon $W_h^* \in W^* = \{W_h^*\}_{h=1}^H$, the aggregate z_h^α for agent α is defined as

$$z_h^\alpha = \int_0^1 W_h^*(\alpha,\beta)\mathcal{L}(s_h^\beta)\,\mathrm{d}\beta, \tag{1}$$

where $\mathcal{L}(s) \in \Delta(\mathcal{S})$ denotes the law of the random variable s. We note that the agents in this game are *heterogeneous*. This means that each agent is affected differently by other agents or, in other words, the aggregates z_h^α for different $\alpha \in \mathcal{I}$ are, in general, different. Given the state $s_h^\alpha \in \mathcal{S}$ and the action $a_h^\alpha \in \mathcal{A}$ of agent α, the agent transitions to a new state $s_{h+1}^\alpha \sim P_h^*(\cdot \,|\, s_h^\alpha, a_h^\alpha, z_h^\alpha)$. The reward functions $r^* = \{r_h^*\}_{h=1}^H$ are deterministic functions $r_h^* : \mathcal{S} \times \mathcal{A} \times \mathcal{M}(\mathcal{S}) \to \mathbb{R}$ for all $h \in [H]$. For agent $\alpha \in \mathcal{I}$ at time h, taking the action a_h^α under the state s_h^α and the aggregate z_h^α earns the agent a reward of $r_h^*(s_h^\alpha, a_h^\alpha, z_h^\alpha)$.

We remark that the above GMFG subsumes the MFG (Xie et al., 2021; Anahtarci et al., 2022) as a special case. To see this, let $W_h(\alpha,\beta) = 1$ for all $\alpha, \beta \in \mathcal{I}$ and $h \in [H]$, then the agents are *homogeneous*. The aggregate z_h^α in Eqn. (1) is simply the state distributions of these homogeneous agents.

A *Markov policy* for the agent $\alpha \in \mathcal{I}$ is characterized by $\pi^\alpha = \{\pi_h^\alpha\}_{h=1}^H \in \Pi^H$, where $\pi_h^\alpha : \mathcal{S} \to \Delta(\mathcal{A})$ lies in the class $\Pi = \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$. The collection of policies of all agents is denoted as $\pi^{\mathcal{I}} = (\pi^\alpha)_{\alpha\in\mathcal{I}} \in \Pi^{\mathcal{I}\times H} = \tilde{\Pi}$. We let $\mu_h^\alpha = \mathcal{L}(s_h^\alpha) \in \Delta(\mathcal{S})$ be the state distribution of the agent α at time h. Then $\mu_h^{\mathcal{I}} = (\mu_h^\alpha)_{\alpha\in\mathcal{I}} \in \Delta(\mathcal{S})^{\mathcal{I}}$ is the set of state distributions of all agents at time h. Note that the aggregate z_h^α is a function of the distributions $\mu_h^{\mathcal{I}}$ and the graphon W_h^*, so we may write it more explicitly as $z_h^\alpha(\mu_h^{\mathcal{I}}, W_h)$. The distribution flow $\mu^{\mathcal{I}} = (\mu_h^{\mathcal{I}})_{h=1}^H \in \Delta(\mathcal{S})^{\mathcal{I}\times H} = \tilde{\Delta}$ consists of the state distributions of all agents at any given time.

In this work, we focus on the regularized problem (Nachum et al., 2017; Cui and Koeppl, 2021a). This setting augments standard reward functions with the entropy of the implemented policy. Some recent works have shown that entropy regularization can accelerate the convergence of the policy gradient methods (Shani et al., 2020; Cen et al., 2022). In a λ-regularized GMFG, when agent α implements policy π_h^α at time h, she will receive a reward $r_h^*(s_h^\alpha, a_h^\alpha, z_h^\alpha) - \lambda \log \pi_h^\alpha(a_h^\alpha \,|\, s_h^\alpha)$ by taking action a_h^α at state s_h^α. Given the underlying distribution flow $\mu^{\mathcal{I}}$ and the policy $\pi^{\mathcal{I}}$, the *value function* and the *action-value function* for agent $\alpha \in \mathcal{I}$ in the λ-regularized game with $\lambda > 0$ are respectively defined as

$$V_m^{\lambda,\alpha}(s, \pi^\alpha, \mu^{\mathcal{I}}, W^*) = \mathbb{E}^{\pi^\alpha}\Bigg[\sum_{h=m}^H r_h^*\big(s_h^\alpha, a_h^\alpha, z_h^\alpha(\mu_h^{\mathcal{I}}, W_h^*)\big) - \lambda \log \pi_h^\alpha(a_h^\alpha \,|\, s_h^\alpha)\,\Bigg|\, s_h^\alpha = s\Bigg],$$

$$Q_h^{\lambda,\alpha}(s, a, \pi^\alpha, \mu^{\mathcal{I}}, W^*) = r_h^*\big(s, a, z_h^\alpha(\mu_h^{\mathcal{I}}, W_h^*)\big) + \mathbb{E}\big[V_{h+1}^{\lambda,\alpha}(s_{h+1}^\alpha, \pi^\alpha, \mu^{\mathcal{I}}, W^*) \,|\, s_h^\alpha = s, a_h^\alpha = a\big],$$

where the expectation $\mathbb{E}^{\pi^\alpha}[\cdot]$ is taken with respect to $a_h^\alpha \sim \pi_h^\alpha(\cdot \,|\, s_h^\alpha)$ and $s_{t+1}^\alpha \sim P_h(\cdot \,|\, s_h^\alpha, a_h^\alpha, z_h^\alpha)$ for all $h \in [H]$. The cumulative reward of agent $\alpha \in \mathcal{I}$ under policy $\pi^{\mathcal{I}}$ is defined as $J^{\lambda,\alpha}(\pi^\alpha, \mu^{\mathcal{I}}, W^*) = \mathbb{E}_{\mu_1^\alpha}[V_1^{\lambda,\alpha}(s, \pi^\alpha, \mu^{\mathcal{I}}, W^*)]$, where the expectation is taken with respect to $s \sim \mu_1^\alpha$.

Definition 1 *A NE of the λ-regularized GMFG is a pair $(\pi^{*,\mathcal{I}}, \mu^{*,\mathcal{I}}) \in \tilde{\Pi} \times \tilde{\Delta}$ that satisfies the following two conditions:*

- *(Agent rationality) $J^{\lambda,\alpha}(\pi^{*,\alpha}, \mu^{*,\mathcal{I}}, W^*) \geq J^{\lambda,\alpha}(\tilde{\pi}^\alpha, \mu^{*,\mathcal{I}}, W^*)$ for all $\alpha \in \mathcal{I}$ and $\tilde{\pi}^\alpha \in \Pi^H$.*

- *(Distribution consistency) The distribution flow $\mu^{*,\mathcal{I}}$ is equal to the distribution flow $\mu^{\pi^{*,\mathcal{I}},\mathcal{I}}$ induced by the policy $\pi^{*,\mathcal{I}}$.*

We define the operator that returns the optimal policy when the underlying distribution flow is $\mu^{\mathcal{I}}$ and the graphon is W as $\Gamma_1^{\lambda}(\mu^{\mathcal{I}}, W) \in \tilde{\Pi}$, i.e., $\pi^{\mathcal{I}} = \Gamma_1^{\lambda}(\mu^{\mathcal{I}}, W)$ if $J^{\lambda,\alpha}(\pi^{\alpha}, \mu^{\mathcal{I}}, W) = \sup_{\tilde{\pi}^{\alpha}\in\tilde{\Pi}} J^{\lambda,\alpha}(\tilde{\pi}^{\mathcal{I}}, \mu^{\mathcal{I}}, W)$ for all $\alpha \in \mathcal{I}$. In this work, we focus on the case where the GMFG is regularized, i.e., $\lambda > 0$. Thus, Γ_1^{λ} is uniquely defined. We also define the operator that returns the distribution flow induced by the policy $\pi^{\mathcal{I}}$ as $\Gamma_2(\pi^{\mathcal{I}}, W^*) \in \tilde{\Delta}$, i.e., $\tilde{\mu}^{\mathcal{I}} = \Gamma_2(\pi^{\mathcal{I}}, W^*)$ if

$$\tilde{\mu}_{h+1}^{\alpha}(s') = \int_{\mathcal{S}} \sum_{a\in\mathcal{A}} \tilde{\mu}_h^{\alpha}(s)\pi_h^{\alpha}(a \mid s) P_h\big(s' \mid s, a, z_h^{\alpha}(\tilde{\mu}_h^{\mathcal{I}}, W_h^*)\big)\mathrm{d}s \qquad \text{for all } s' \in \mathcal{S}, h \in [H-1] \text{ and } \alpha \in \mathcal{I},$$

and $\tilde{\mu}_1^{\mathcal{I}} = \mu_1^{\mathcal{I}}$. Our goal in this paper is to learn the NE of the λ-regularized GMFG from the data collected of the sampled agents. Before giving an overview of our agent sampling schemes, we first introduce the equivariance property of GMFG.

3.1 Equivariance Property of GMFGs

We now argue that GMFG is *equivariant* to the measure-preserving bijection imposed on agents. In the GMFG, all the interactions among agents are captured by the underlying graphons. For agents $\alpha, \beta \in \mathcal{I}$, the value $W(\alpha, \beta)$ represents the strength of interactions between α and β. Intuitively, if we "permute" the positions of agents in the graphon (i.e., we "permute" the values of $\alpha \in \mathcal{I}$) and transform the graphons accordingly, the resultant game remains the same up to this permutation. However, given an uncountable number of agents in $[0, 1]$, the concept of "permutation" of finite objects should be more precisely stated. This is formalized by the notion of *measure-preserving bijections* from $[0, 1]$ to $[0, 1]$. Given a measure-preserving bijection $\phi : [0, 1] \to [0, 1]$, the transformation of a graphon W^{ϕ} is defined as

$$W^{\phi}(x, y) = W\big(\phi(x), \phi(y)\big).$$

We denote the set of all the measure-preserving bijections as $\mathcal{B}_{[0,1]}$. Then the equivariance property of the GMFG can be stated as follows.

Proposition 2 *For any policy $\pi^{\mathcal{I}} \in \tilde{\Pi}$, let its distribution flow on $(\mathcal{S}, \mathcal{A}, \mu_1, H, P^*, r^*, W^*)$ be $\mu^{\mathcal{I}} \in \tilde{\Delta}$. In other words, $\mu^{\mathcal{I}} = \Gamma_2(\pi^{\mathcal{I}}, W^*)$. For any $\phi \in \mathcal{B}_{[0,1]}$, define the ϕ-transformed policy $\pi^{\phi,\mathcal{I}}$ as $\pi^{\phi,\alpha} = \pi^{\phi(\alpha)}$ for all $\alpha \in \mathcal{I}$. Then we denote its distribution flow on $(\mathcal{S}, \mathcal{A}, \mu_1, H, P^*, r^*, W^{\phi,*})$ as $\mu^{\phi,\mathcal{I}} \in \tilde{\Delta}$, i.e., $\mu^{\phi,\mathcal{I}} = \Gamma_2(\pi^{\phi,\mathcal{I}}, W^{\phi,*})$. We have*

$$\mu^{\phi,\alpha} = \mu^{\phi(\alpha)} \text{ for all } \alpha \in \mathcal{I}.$$

The proof is provided in Appendix D. Proposition 2 shows that the graphons transformed by a measure-preserving bijections defines the same game as the original graphons up to the bijection. In Section 5.5, we learn the values of the graphon from the sampled agents without information of their positions. This proposition shows that we can learn the graphon up to a measure-preserving bijection, which motivates the definition of the permutation-invariance risk in Section 5.5.

3.2 Overview of Sampling Schemes

Our goal is to design algorithms for a *central planner* to learn the NE from the data collected from a subset of sampled agents in a simulator. This simulator setting is widely accepted and adopted in the MFG community (Guo et al., 2019; Anahtarci et al., 2022). We further note that MFGs constitute a subclass of GMFGs. When all the agents indexed by $[0, 1]$ implement the behavior policies, the simulator will sample N agents from $[0, 1]$ and collect their states, actions, and rewards. The learner (central planner) only has access to the samples of the sampled agents. In this work, we consider three types of agent sampling procedures

1. Agents are sampled from *known grid* positions. In particular, we sample the agents at grid positions $\{i/N\}_{i=1}^{N} \subset [0, 1]$, and we know the position of each agent;
2. Agents are sampled from *known random* positions. In particular, we sample the agents from N i.i.d. samples of $\mathrm{Unif}([0, 1])$, and the positions of agents are also known;
3. Agents are sampled from *grid* positions, but the positions of the sampled agents are *unknown*. For example, we know the positions of the sampled agents belong to the set $\{i/N\}_{i=1}^{N}$. However, the position of each agent within the set $\{i/N\}_{i=1}^{N}$ is unknown.

In Section 5, we design and analyze a model learning algorithm that estimates the transition kernel P^*, the reward function r^*, and the underlying graphons W^* for each of these three sampling schemes. In the first two cases, we design a model learning algorithm that estimates the transition kernel P^*, the reward function r^*, and the underlying graphons W^*. However, in the third case, we cannot estimate the original graphons, since the positions of the agents are unknown. Instead, we can only estimate the original graphons up to a measure-preserving bijection. In this case, we need to recover the "relative positions" of sampled agents to select the graphons from set $\tilde{\mathcal{W}}$. For N agents, there are $N!$ potential cases for their relative positions. The super-exponential size of the search space makes the problem statistically challenging. To complete the story, there is a sampling scheme where the positions of agents are unknown and random. However, the analysis of algorithms in this case is difficult due to the need to carefully analyze the order statistics which is rather different from the abovementioned three cases. We leave this case for future work.

4. Learning Algorithm for GMFG

4.1 Design of the GMFG-PPO Algorithm

In this section, we design an algorithm called GMFG-PPO (Algorithm 1) to learn an NE of the λ-regularized GMFG with $\lambda > 0$. GMFG-PPO, which is an iterative algorithm, involves three main steps in each iteration. First, it evaluates the distribution flow and the action-value function (Line 4), assuming the access to a sub-module for computing these. In Section 5, we design this sub-module as a model-based learning algorithm. Second, it updates the distribution flow as a mixture of the distribution flow estimate $\hat{\mu}_t^{\mathcal{I}}$ of the current policy $\pi_t^{\mathcal{I}}$ and the current distribution flow $\hat{\bar{\mu}}_t^{\mathcal{I}}$ (Line 5). The "bar" notation (such as $\bar{\mu}^{\mathcal{I}}$) represents a mixture of distributions, and the "hat" notation (such as $\hat{\mu}_t^{\mathcal{I}}$) represents estimated distributions. All procedures are repeated T times in the algorithm. This

Algorithm 1 GMFG-PPO

Procedure:

1: Initialize $\pi^{\alpha}_{1,h}(\cdot\,|\,s) = \mathrm{Unif}(\mathcal{A})$ for all $s \in \mathcal{S}$, $h \in [H]$ and $\alpha \in \mathcal{I}$.

2: Initialize $\hat{\bar{\mu}}^{\mathcal{I}}_1 = \hat{\Gamma}_2(\pi^{\mathcal{I}}_1, \hat{W})$.

3: **for** $t = 1, 2, \ldots, T$ **do**

4: Compute the distribution flow $\hat{\mu}^{\mathcal{I}}_t = \hat{\Gamma}_2(\pi^{\mathcal{I}}_t, \hat{W})$ induced by policy $\pi^{\mathcal{I}}_t$ and corresponding action-value function $\hat{Q}^{\lambda,\alpha}_h(s, a, \pi^{\alpha}_t, \hat{\bar{\mu}}^{\mathcal{I}}_t, \hat{W})$ for all $\alpha \in \mathcal{I}$ and $h \in [H]$.

5: $\hat{\bar{\mu}}^{\mathcal{I}}_{t+1} = (1-\alpha_t)\hat{\bar{\mu}}^{\mathcal{I}}_t + \alpha_t \hat{\mu}^{\mathcal{I}}_t$.

6: $\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s) \propto \big(\pi^{\alpha}_{t,h}(\cdot\,|\,s)\big)^{1-\frac{\lambda\eta_{t+1}}{1+\lambda\eta_{t+1}}} \exp\Big(\frac{\eta_{t+1}}{1+\lambda\eta_{t+1}} \hat{Q}^{\lambda,\alpha}_h(s, \cdot, \pi^{\alpha}_t, \hat{\bar{\mu}}^{\mathcal{I}}_t, \hat{W})\Big)$ for all $\alpha \in \mathcal{I}$ and $h \in [H]$

7: $\pi^{\alpha}_{t+1,h}(\cdot\,|\,s) = (1-\beta_{t+1})\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s) + \beta_{t+1}\mathrm{Unif}(\mathcal{A})$

8: **end for**

9: Output $\bar{\pi}^{\mathcal{I}} = \mathrm{Unif}\big(\pi^{\mathcal{I}}_{[1:T]}\big)$ and $\bar{\mu}^{\mathcal{I}} = \mathrm{Unif}\big(\hat{\bar{\mu}}^{\mathcal{I}}_{[1:T]}\big)$

procedure is known as *fictitious play* in Xie et al. (2021) and Perrin et al. (2020). It slows down the update of the distribution flow. In our analysis, this deceleration is shown to be important for learning the optimal policy with respect to the current distribution flow. Finally, we improve the policy with one-step mirror descent (Line 6). We note that Line 6 is in fact the closed-form solution to the optimization

$$\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s) = \operatorname*{argmax}_{p \in \Delta(\mathcal{A})} \eta_{t+1}\Big[\big\langle \hat{Q}^{\lambda,\alpha}_h(s, \cdot, \pi^{\alpha}_t, \hat{\bar{\mu}}^{\mathcal{I}}_t, \hat{W}), p\big\rangle - \lambda \bar{H}(p)\Big] - \mathrm{KL}\big(p \| \pi^{\alpha}_{t,h}(\cdot\,|\,s)\big) \quad \forall\, s \in \mathcal{S},$$

where $\bar{H}(p) = \langle p, \log p\rangle$ is the negative entropy function. This procedure is one-step policy mirror descent in Lan (2022), and it also corresponds to the PPO algorithm in Schulman et al. (2017). This policy improvement procedure aims to optimize the policy in the MDP induced by $\hat{\bar{\mu}}^{\mathcal{I}}_t$. With the convergence of $\hat{\bar{\mu}}^{\mathcal{I}}_t$ to $\mu^{*,\mathcal{I}}$, this procedure can learn the optimal policy on $\mu^{*,\mathcal{I}}$, i.e., the policy $\pi^{*,\mathcal{I}}$ in the NE. Line 7 mixes the current policy iterate $\hat{\pi}^{\mathcal{I}}$ with the uniform distribution. Intuitively, this mixing ensures that the policy has sufficient exploration in order to find the NE eventually.

GMFG-PPO differs from the NE learning algorithm of regularized MFG in Xie et al. (2021) in three aspects. First, GMFG-PPO is designed to learn the NE of the regularized GMFG. It involves graphon learning and requires the policy and action-value function updates for all the agents. In contrast, the algorithm in Xie et al. (2021) can only learn the NE of the regularized MFG, which is a special case of GMFG with *constant* graphons. It only keeps track of the policy and action-value function of a representative agent. Second, GMFG-PPO learns a non-stationary NE, whereas the algorithm in Xie et al. (2021) learns a *stationary* NE. Finally, the stepsize η_t used in the policy improvement (Line 6) will be set to be a (non-vanishing) constant in Section 4.2. In contrast, the algorithm in Xie et al. (2021) sets $\eta_t = o(1)$. Our choice of η_t is the chief reason for the improved convergence rate.

4.2 Convergence Analysis of GMFG-PPO

Assuming that an NE exists (Cui and Koeppl, 2021b; Fabian et al., 2022), we now present convergence results for learning it. We denote an NE of the λ-regularized GMFG as

$(\pi^{*,\mathcal{I}}, \mu^{*,\mathcal{I}})$. We measure the distances between policies and distribution flows with

$$D(\pi^{\mathcal{I}}, \tilde{\pi}^{\mathcal{I}}) = \int_0^1 \sum_{h=1}^{H} \mathbb{E}_{\mu_h^{*,\alpha}}\Big[\big\|\pi_h^\alpha(\cdot \,|\, s) - \tilde{\pi}_h^\alpha(\cdot \,|\, s)\big\|_1\Big]\, \mathrm{d}\alpha, \qquad \text{and}$$

$$d(\mu^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}}) = \int_0^1 \sum_{h=1}^{H} \|\mu_h^\alpha - \tilde{\mu}_h^\alpha\|_1\, \mathrm{d}\alpha.$$

For the purpose of our convergence results, we make a few assumptions about the λ-regularized GMFG. We first assume the Lipschitz continuity of transition kernels and reward functions.

Assumption 1 *The reward function $r_h(s,a,z)$ is Lipschitz continuous in z for all $h \in [H]$, that is $|r_h(s,a,z) - r_h(s,a,z')| \leq L_r\|z - z'\|_1$ for all $h \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The transition kernel $P_h(\cdot \,|\, s,a,z)$ is Lipschitz continuous in z with respect to the total variation, that is $\mathrm{TV}(P_h(\cdot \,|\, s,a,z), P_h(\cdot \,|\, s,a,z')) \leq L_P\|z - z'\|_1$ for all $h \in [H]$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

This assumption is common in the MFG and GMFG literature (Cui and Koeppl, 2021b; Anahtarci et al., 2022). We then assume that the composition of the operators Γ_1^λ and Γ_2 is contractive in the following sense.

Assumption 2 *There exist constants $d_1, d_2 > 0$ and $d_1 d_2 < 1$ such that for any policies $\pi^{\mathcal{I}}, \tilde{\pi}^{\mathcal{I}}$ and distribution flows $\mu^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}}$, it holds that*

$$D\big(\Gamma_1^\lambda(\mu^{\mathcal{I}}, W^*), \Gamma_1^\lambda(\tilde{\mu}^{\mathcal{I}}, W^*)\big) \leq d_1\, d(\mu^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}}), \quad \textit{and}$$

$$d\big(\Gamma_2(\pi^{\mathcal{I}}, W^*), \Gamma_2(\tilde{\pi}^{\mathcal{I}}, W^*)\big) \leq d_2\, D(\pi^{\mathcal{I}}, \tilde{\pi}^{\mathcal{I}}).$$

This "contractive" assumption plays an important role in the design of efficient algorithms, since it guarantees the convergence of both $\pi^{\mathcal{I}}$ and $\mu^{\mathcal{I}}$ using simple fixed point iterations. This assumption is widely adopted in the MFG literature (Xie et al., 2021; Guo et al., 2019), and it holds if the regularization λ is higher enough than L_r and L_P (Anahtarci et al., 2022; Cui and Koeppl, 2021a). We note that Assumption 2 indeed implies the existence and uniqueness of NE.

Proposition 3 *Under Assumption 2, the λ-regularized GMFG admits exactly one NE up to a set of zero-measure agents with respect to the Lebesgue measure on $[0,1]$.*

The proof is provided in Appendix O. For a policy $\pi^{\mathcal{I}}$ and any distribution flow $\mu^{\mathcal{I}}$, we define the operator Γ_3 that satisfies $\mu^{+,\mathcal{I}} = \Gamma_3(\pi^{\mathcal{I}}, \mu^{\mathcal{I}}, W)$ as

$$\mu_1^{+,\mathcal{I}} = \mu_1^{\mathcal{I}}, \qquad \mu_{h+1}^{+,\alpha}(s') = \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} \mu_h^{+,\alpha}(s)\pi_h^\alpha(a \,|\, s) P_h\big(s' \,|\, s, a, z_h^\alpha(\mu_h^{\mathcal{I}}, W_h)\big)\, \mathrm{d}s,$$

for all $s' \in \mathcal{S}, \alpha \in \mathcal{I}$, and $h \geq 1$. The operator Γ_3 outputs the distribution flow $\mu^{+,\mathcal{I}}$ for implementing the policy $\pi^{\mathcal{I}}$ on the MDP induced by $\mu^{\mathcal{I}}$. We now make an assumption about certain concentrability coefficients.

Assumption 3 *For any distribution flow $\mu^{\mathcal{I}}$, we define its induced optimal policy on the MDP induced by it as $\pi_\mu^{*,\mathcal{I}} = \Gamma_1^\lambda(\mu^{\mathcal{I}}, W^*)$ and the induced distribution flow as $\tilde{\mu}^{*,\mathcal{I}} = \Gamma_3(\pi_\mu^{*,\mathcal{I}}, \mu^{\mathcal{I}}, W^*)$. Then there exists a constant $C_\mu > 0$ such that for any distribution flow $\mu^{\mathcal{I}}$, it hold that*

$$\sup_{\alpha\in\mathcal{I}, h\in[H]} \mathbb{E}_{s\sim\tilde{\mu}_h^{*,\alpha}}\left[\left|\frac{\mu_h^{*,\alpha}(s)}{\tilde{\mu}_h^{*,\alpha}(s)}\right|^2\right] \le C_\mu^2.$$

This assumption concerns the boundedness of concentrability coefficients. This type of assumption are standard in the policy optimization literatures (Shani et al., 2020; Bhandari and Russo, 2019; Agarwal et al., 2020). The policy $\pi_\mu^{*,\alpha}(a_h|s_h) > 0$ is strictly positive for all $a_h \in \mathcal{A}, s_h \in \mathcal{S}$ due to the presence of the regularization. When $\mathcal{S}$ is finite, this assumption holds if $P_h^*(s_{h+1}|s_h, a_h, z_h) > 0$ for all $s_{h+1}, s_h \in \mathcal{S}$, $a_h \in \mathcal{A}$, and $z_h \in \mathcal{M}(\mathcal{S})$. When $\mathcal{S}$ is compact (but uncountable) this assumption holds if $P_h^*(s_{h+1}|s_h, a_h, z_h) \le \kappa P_h^*(s_{h+1}|s_h, a_h, z_h')$ for some constant $\kappa > 0$ and all $s_{h+1}, s_h \in \mathcal{S}$, $a_h \in \mathcal{A}$, and $z_h, z_h' \in \mathcal{M}(\mathcal{S})$. We then make an assumption about the accuracy about our distribution flow and action-value function estimates in Line 4 of Algorithm 1.

Assumption 4 *We have access to the estimator $\hat{P} = \{\hat{P}_h\}_{h=1}^H$, $\hat{r} = \{\hat{r}_h\}_{h=1}^H$, and $\hat{W} = \{\hat{W}_h\}_{h=1}^H$ and corresponding operator estimate $\hat{\Gamma}_2(\cdot, \hat{W})$ and action-value function estimator $\hat{Q}_h^{\lambda,\alpha}(\cdot, \hat{W})$. These estimates satisfy that for any policy $\pi^{\mathcal{I}}$, we have that*

$$d\big(\hat{\Gamma}_2(\pi^{\mathcal{I}}, \hat{W}), \Gamma_2(\pi^{\mathcal{I}}, W^*)\big) \le \varepsilon_\mu,$$

and that for any policy $\pi^{\mathcal{I}}$ and distribution flow $\mu^{\mathcal{I}}$

$$\sup_{\tilde{\pi}^{\mathcal{I}},\alpha} \mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\big\|\hat{Q}_h^{\lambda,\alpha}(s,\cdot,\pi^\alpha,\mu^{\mathcal{I}},\hat{W}) - Q_h^{\lambda,\alpha}(s,\cdot,\pi^\alpha,\mu^{\mathcal{I}},W^*)\big\|_\infty \le \varepsilon_Q.$$

for some constants ε_μ and ε_Q.

We make this assumption only for ease of the presentation of the analysis of our algorithm. In Section 6, we will replace this assumption with the actual performance guarantee of our model learning algorithms. When learning the model from L trajectories of N sampled agents, we could quantify ε_μ and ε_Q as: (i) (known fixed positions) $\varepsilon_\mu = O(N^{-1/2} + (NL)^{-1/4})$ and $\varepsilon_Q = O(N^{-1/2} + (NL)^{-1/2})$. (ii) (known random positions) $\varepsilon_\mu = O(N^{-1/2} + (NL)^{-1/4})$ and $\varepsilon_Q = O(N^{-1/4} + (NL)^{-1/2})$. (iii) (unknown fixed positions) $\varepsilon_\mu = O(N^{-1/2} + (N/L)^{1/4})$ and $\varepsilon_Q = O(N^{-1/2} + (L)^{-1/2})$.

Theorem 4 *We set $\alpha_t = O(T^{-2/3})$, $\beta_t = O(T^{-1})$, and η_t to a constant that only depends on λ, H and $|\mathcal{A}|$. Under Assumptions 1, 2, 3, and 4, Algorithm 1 returns the policy $\bar{\pi}^{\mathcal{I}}$ and the distribution flow $\bar{\mu}^{\mathcal{I}}$ that satisfies*

$$D\bigg(\frac{1}{T}\sum_{t=1}^T \pi_t^{\mathcal{I}}, \pi^{*,\mathcal{I}}\bigg) + d\bigg(\frac{1}{T}\sum_{t=1}^T \hat{\bar{\mu}}_t^{\mathcal{I}}, \mu^{*,\mathcal{I}}\bigg) = O\bigg(\frac{\sqrt{\log T}}{T^{1/3}}\bigg) + O(\varepsilon_\mu + \sqrt{\varepsilon_Q + \varepsilon_\mu}).$$

The proof is provided in Appendix E. There are two main differences in Theorem 4 and Xie et al. (2021, Theorem 1). First, we achieve a faster rate $\tilde{O}(T^{-1/3})$ than the rate $\tilde{O}(T^{-1/5})$ in Xie et al. (2021). This improvement is attributed to the newly designed stepsize η_t, which is a *constant*, but the algorithm in Xie et al. (2021) sets η_t to be $O(T^{-2/5})$. Intuitively, a stepsize η_t that is independent of T will result in faster convergence of an algorithm compared to one that decays as T grows. However, the proof involves a novel optimization error recursion analysis for this new stepsize. This novel optimization error recursion analysis also generalizes Lan (2022, Theorem 1) to the time-inhomogeneous MDP with a finite horizon. See Appendix F for the statement. Second, Theorem 4 does not require the first condition in Assumptions 4 and 5 in Xie et al. (2021). Instead, we adopt the more realistic Assumption 1 concerning the Lipschitzness of transition kernels and reward functions to control the difference between the MDP induced by difference distribution flows.

5. Model Estimation From Datasets

We assume that the state space $\mathcal{S} \subseteq \mathbb{R}^{d_\mathrm{s}}$ is a subset of $\mathbb{R}$, i.e., $d_\mathrm{s} = 1$. Our results can be extended to the case $d_\mathrm{s} > 1$ by using kernels of functions with multiple outputs. Since $\mathcal{S}$ is compact, there exists a constant $B_S > 0$ such that $|s| \leq B_S$ for all $s \in \mathcal{S}$.

5.1 Dataset Collection

Since the GMFG involves uncountably infinite agents, it is impossible to collect the trajectories of all the agents. Thus, we sample N agents $\{\xi_i\}_{i=1}^N$ in $[0,1]$ to collect their states, actions, and rewards in each episode. We consider three sampling methods: (i) agents' positions $\{\xi_i\}_{i=1}^N$ are known grids, namely, $\xi_i = i/N$ for all $i \in [N]$. Furthermore, the map between the identity of each agent to the grid $\{i/N\}_{i=1}^N$ is known. (ii) $\{\xi_i\}_{i=1}^N$ are known i.i.d. samples of the uniform distribution $\mathrm{Unif}([0,1])$. (iii) agents' positions $\{\xi_i\}_{i=1}^N$ are grid points, and these positions are *unknown*. Then we acquire the states and actions of these sampled agents. For notational simplicity, we denote the state $s_h^{\xi_i}$ and action $a_h^{\xi_i}$ of the agent ξ_i as $s_h^i = s_h^{\xi_i}$ and $a_h^i = a_h^{\xi_i}$, respectively. To collect these data, we implement L behavior policies $\pi_\tau^{\mathcal{I}}$ for all $\tau \in [L]$. In the τ^{th} episode, a trajectory of these agents is $\mathcal{D}_\tau = \{(s_{\tau,h}^{[N]}, a_{\tau,h}^{[N]}, r_{\tau,h}^{[N]}, s_{\tau,h+1}^{[N]})\}_{h=1}^H$. The dataset consists of L trajectories, i.e., $\mathcal{D} = \{\mathcal{D}_\tau\}_{\tau=1}^L$.

We note that once the behavior policy $\pi_\tau^{\mathcal{I}}$ is determined, the distribution flow $\mu_\tau^{\mathcal{I}}$ is fixed. Then the influence aggregate on the i^{th} agent $z_{\tau,h}^i(W_h^*)$ is a function only of ξ_i, which is independent of the states of other agents. Thus, the distribution of $s_{\tau,h}^{[N]}$ is $\prod_{i=1}^N \mu_{\tau,h}^{\xi_i} = \prod_{i=1}^N \mu_{\tau,h}^i$.

5.2 Mean-Embedding of Distribution Flows

The transition kernels and the reward functions both take (s, a, z) as their inputs. However, the aggregate $z \in \mathcal{M}(\mathcal{S})$ for an agent which is defined in Eqn. (1) is not available to us, since it requires the unknown values of graphons W^* and the distribution flow $\mu^{\mathcal{I}}$. From the collected data, we only have the states $\{s_{\tau,h}^i\}_{i=1}^N$ sampled from distributions $\{\mu_{\tau,h}^i\}_{i=1}^N$. Thus, we first need to estimate the distribution flow $\mu^{\mathcal{I}}$ from these sample. We handle this by using a mean-embedding, which is a widely adopted method in distribution regression (Szabó et al., 2016, 2015). Define $\Xi = \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, then $d_{s,a,z} = \delta_s \times \delta_a \times z$ is measure on Ξ. Given a positive definite kernel $k : \Xi\times\Xi \to \mathbb{R}$, we denote the Reproducing Kernel Hilbert Space (RKHS) spanned by kernel k as $\mathcal{H}$. Then we embed the measure $d_{s,a,z}$ with the kernel k as

$$\omega_{d_{s,a,z}} = \int_{\mathcal{S}} k\big(\cdot,(s,a,s')\big) z(\mathrm{d}s').$$

We have $\omega_{d_{s,a,z}} \in \mathcal{H}$. We note that such mean-embedding procedure will not cause the problem to be degenerate, since the embedding with the identity kernel degenerates to $d_{s,a,z}$. For our regression setting, we will embed the measure $\delta_{s_h^\alpha}\times\delta_{a_h^\alpha}\times z_h^\alpha(W_h^*)$ for all $\alpha\in\mathcal{I}$, and $h\in[H]$. Here the aggregate z_h^α is the influence aggregate for agent α at time h defined in Eqn. (1). Then the mean-embedding of the measure $\delta_{s_h^\alpha}\times\delta_{a_h^\alpha}\times z_h^\alpha(W_h^*)$ is

$$\omega_h^\alpha(W_h^*) = \int_0^1\int_{\mathcal{S}} W_h^*(\alpha,\beta) k\big(\cdot,(s_h^\alpha,a_h^\alpha,s)\big)\mu_h^\beta(s)\,\mathrm{d}s\,\mathrm{d}\beta.$$

Given such embedding representation, we reformulate the transition kernels and the reward functions as functions $f_h^*, g_h^* : \mathcal{H}\to\mathbb{R}$ that is defined as

$$s_{h+1}^\alpha = f_h^*\big(\omega_h^\alpha(W_h^*)\big) + \varepsilon_h, \quad r_h^\alpha = g_h^*\big(\omega_h^\alpha(W_h^*)\big) \text{ for all } h\in[H], \alpha\in\mathcal{I}, \tag{2}$$

where $\{\varepsilon_h^\alpha\}_{\alpha\in\mathcal{I}}$ are independent zero-mean noises. Since $|s|\le B_S$, we have $|\varepsilon_h^\alpha|\le 2B_S$.

5.3 Assumptions for Model Learning

In the following, we will estimate the transition kernels $\{f_h^*\}_{h=1}^H$, the reward functions $\{g_h^*\}_{h=1}^H$ and the graphons $\{W_h^*\}_{h=1}^H$ from the collected data. With nonparametric regression methods, we adopt a general graphon class $\tilde{\mathcal{W}}$ to estimate the underlying graphons and adopt the kernels $\bar{K} : \mathcal{H}\times\mathcal{H}\to\mathbb{R}$ and $\tilde{K} : \mathcal{H}\times\mathcal{H}\to\mathbb{R}$ to estimate the transition kernel and reward functions, respectively. The space spanned by the kernels $\bar{K}$ and $\tilde{K}$ are respectively denoted as $\bar{\mathcal{H}}$ and $\tilde{\mathcal{H}}$. We postpone the details of the estimation algorithms for three sampling schemes to the following sections, and we first state the assumptions needed for the convergence of all these estimation algorithms.

First, we assume the Lipschitz continuity of the graphon class $\tilde{\mathcal{W}}$ and the nominal graphons $W^* = \{W_h^*\}_{h=1}^H$. This assumption will help us to generalize the estimate from the sampled agents to the unobserved agents.

Assumption 5 (Lipschitzness of Graphons) *For any $\tilde{W}\in\tilde{\mathcal{W}}$ (resp. $\{W_h^*\}_{h=1}^H$), we have that $|\tilde{W}(\alpha,\beta)-\tilde{W}(\alpha',\beta')|\le L_{\tilde{\mathcal{W}}}(|\alpha-\alpha'|+|\beta-\beta'|)$ (resp. $|W(\alpha,\beta)-W(\alpha',\beta')|\le L_{W^*}(|\alpha-\alpha'|+|\beta-\beta'|)$) for all $\alpha,\alpha',\beta,\beta'\in[0,1]$, where $L_{\tilde{\mathcal{W}}}>0$ (resp. $L_{W^*}>0$) is a constant.*

For ease of notation, we define $L_{\mathcal{W}} = \max\{L_{\tilde{\mathcal{W}}}, L_{W^*}\}$. Second, we assume the boundedness and the Lipschitz continuity of the kernels. Similar as Assumption 5, this assumption is helpful to guarantee the boundedness of estimates and generalize the estimates from the sampled agents to the unobserved agents.

Assumption 6 (Boundedness and Lipschitzness of Kernels) *The reproducing kernels k, $\bar{K}$ and $\tilde{K}$ satisfy*

- *The kernel k is bounded, i.e., there exists $B_k > 0$ such that $k(x, x) \leq B_k^2$ for all $x \in \Xi$.*
- *The kernel $\bar{K}$ (resp. $\tilde{K}$) is bounded, i.e., there exists $B_{\bar{K}} > 0$ (resp. $B_{\tilde{K}} > 0$) such that $\bar{K}(\omega, \omega) \leq B_{\bar{K}}^2$ (resp. $\tilde{K}(\omega, \omega) \leq B_{\tilde{K}}^2$) for all $\omega \in \mathcal{H}$.*
- *The kernel $\bar{K}$ (resp. $\tilde{K}$) is $L_{\bar{K}}$-Lipschitz (resp. $L_{\tilde{K}}$) continuous, i.e., $\|\bar{K}(\cdot, \omega) - \bar{K}(\cdot, \omega')\|_{\bar{\mathcal{H}}} \leq L_{\bar{K}} \|\omega - \omega'\|_{\mathcal{H}}$ (resp. $\|\tilde{K}(\cdot, \omega) - \tilde{K}(\cdot, \omega')\|_{\tilde{\mathcal{H}}} \leq L_{\tilde{K}} \|\omega - \omega'\|_{\mathcal{H}}$) for all $\omega, \omega' \in \mathcal{H}$.*

For ease of notation, we define the maximal boundedness parameter $B_K = \max\{B_{\bar{K}}, B_{\tilde{K}}\}$ and the maximal Lipschitz constant $L_K = \max\{L_{\bar{K}}, L_{\tilde{K}}\}$. Finally, we state the realizability assumption. It guarantees that we choose the proper function class for our regression task. We define the r-ball in a RKHS $\bar{\mathcal{H}}$ as $\mathbb{B}(r, \bar{\mathcal{H}}) = \{f \in \bar{\mathcal{H}} \mid \|f\|_{\bar{\mathcal{H}}} \leq r\}$.

Assumption 7 (Realizability) *The nominal transition functions f_h^*, reward functions g_h^* and graphons W_h^* satisfy that $f_h^* \in \mathbb{B}(r, \bar{\mathcal{H}})$, $g_h^* \in \mathbb{B}(\tilde{r}, \tilde{\mathcal{H}})$ and $W_h^* \in \tilde{\mathcal{W}}$ for all $h \in [H]$, where $r, \tilde{r} > 0$ are some constants.*

For ease of notation, we define the maximal radius as $\bar{r} = \max\{r, \tilde{r}\}$. We note that our algorithms and analysis are also applicable to the general function class $\mathcal{F}$ and $\tilde{\mathcal{F}}$, replacing $\mathcal{H}$ and $\tilde{\mathcal{H}}$. This assumption is realized when the chosen function classes are large enough. For example, $\mathcal{H}$ and $\tilde{\mathcal{H}}$ can be chosen as kernels spaces of neural networks (Jacot et al., 2018), and $\tilde{\mathcal{W}}$ can be a set of neural networks for the purpose of graphon estimation (Xia et al., 2023). In addition to these non-parameteric function classes, for the case in which we know the form of the underlying graphon (e.g., $W_h^*(\alpha, \beta) = a - b \cdot (\alpha + \beta)$ for some $a, b \in \mathbb{R}$), we can also choose the graphon class accordingly (e.g., $\tilde{\mathcal{W}} = \{W \mid W(\alpha, \beta) = k_1 - k_2 \cdot (\alpha + \beta) \text{ for } k_1, k_2 \in \mathbb{R}\}$). Here we adopt the RKHS for $\bar{\mathcal{H}}$ and $\tilde{\mathcal{H}}$ the ease of representation.

5.4 Learning from Sampled Agents with Known Positions

In this section, we design regression algorithms when the positions of sampled agents are known. From the data collection procedure in Section 5.1, the values of the distribution flows $\mu_\tau^{\mathcal{I}}$ for $\tau \in [L]$ are not directly accessible. For the i^{th} agent, the mean-embedding of her state, action and the aggregate at time h in the τ^{th} episode is

$$\omega_{\tau,h}^i(W_h^*) = \int_{\mathcal{S}} k\big(\cdot, (s_{\tau,h}^i, a_{\tau,h}^i, s)\big) z_{\tau,h}^i(\mathrm{d}s) = \int_0^1 \int_{\mathcal{S}} W_h^*(\xi_i, \beta) k\big(\cdot, (s_{\tau,h}^i, a_{\tau,h}^i, s)\big) \mu_{\tau,h}^\beta(s) \,\mathrm{d}s \,\mathrm{d}\beta. \tag{3}$$

Thus, the input of f_h^* and g_h^*, i.e., $\omega_{\tau,h}^i(W_h^*)$, needs to be estimated. Given any graphon $W_h \in \tilde{\mathcal{W}}$, we derive the empirical estimate of the aggregate of the i^{th} agent at time h as

$$\hat{z}_{\tau,h}^i(W_h) = \frac{1}{N-1} \sum_{j \neq i} W_h(\xi_i, \xi_j) \delta_{s_{\tau,h}^j}.$$

This estimate involves three kinds of error sources. The first is the *graphon estimation error*, which originates from the difference between W_h and W_h^*. The second is the *agent sampling error* which originates from the approximation of uncountably many agents in $[0, 1]$ with

$N-1$ of them, i.e., an integral over $[0,1]$ is replaced by a sum over $N-1$ terms. The last is the *state sampling error* in which we replace the integral of $\mu_{\tau,h}^{\xi_j}$ over state space $\mathcal{S}$ with the singleton $\delta_{s_{\tau,h}^j}$. In the analysis, we handle these three errors separately. Given the aggregate estimate $\hat{z}_{\tau,h}^i(W_h)$, the corresponding mean-embedding of the state, action, and the aggregate for the i^{th} agent is

$$\hat{\omega}_{\tau,h}^i(W_h) = \frac{1}{N-1}\sum_{j\neq i} W_h(\xi_i,\xi_j)k\big(\cdot,(s_{\tau,h}^i, a_{\tau,h}^i, s_{\tau,h}^j)\big). \tag{4}$$

Taking this estimate as the input of f_h^* and g_h^*, we evaluate the square error of the prediction and derive the estimates by minimizing the error. Thus, the estimation procedure for learning the system dynamics, the reward functions, and the graphons can be expressed as

$$(\hat{f}_h, \hat{g}_h, \hat{W}_h) = \underset{\substack{f\in\mathbb{B}(r,\tilde{\mathcal{H}}), g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}),\\ \tilde{W}\in\tilde{\mathcal{W}}}}{\operatorname{argmin}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s_{\tau,h+1}^i - f\big(\hat{\omega}_{\tau,h}^i(\tilde{W})\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\hat{\omega}_{\tau,h}^i(\tilde{W})\big)\Big)^2. \tag{5}$$

We note that the above optimization problem is, in general, non-convex. However, we focus on the statistical property of it in this work, and the practical implementation can be done with the help of non-convex optimization algorithms. In this estimation procedure, we form our predictions of states/rewards via the composition of two procedures, i.e.,

$$\{(s_{\tau,h}^i, a_{\tau,h}^i)\}_{i=1}^N \xrightarrow{k,W} \hat{\omega}_{\tau,h}^i(W) \xrightarrow{f \text{ or } g} s_{\tau,h+1}^i / r_{\tau,h}^i \tag{6}$$

In the first stage, the states and actions are embedded with the kernel k and a selected graphon W. In the second stage, the mean-embedding $\hat{\omega}_{\tau,h}^i(W)$ is forwarded by the functions in $\mathcal{H}$ or $\tilde{\mathcal{H}}$.

This two-stage prediction distinguishes our estimation procedure from the algorithms designed for the distribution regression problem (Szabó et al., 2016; Fang et al., 2020; Meunier et al., 2022). In the distribution regression problem, the covariate, i.e., the input of f or g in Eqn. (6), is an unknown distribution. In this problem, we are tasked with performing a regression from the data of the response variable and the i.i.d. samples of the unknown distribution. Although the distribution regression problem also requires a two-stage prediction similarly as Eqn. (6), i.e., the covariate should be first estimated from i.i.d. samples drawn from itself, our problem setting involving graphons is a strict generalization of distribution regression. First, the input of f or g in our problem is a function of a set of distributions $\{\mu_{\tau,h}^\alpha\}_{\alpha\in\mathcal{I}}$. In contrast, the covariate of the distribution regression problem is a single distribution. Second, in addition to the recovery of $\mu_{\tau,h}^{\mathcal{I}}$ from its samples, our problem requires the estimation of the graphon W to form $\hat{\omega}_{\tau,h}^i(W)$. However, the distribution regression problem only requires the recovery of a distribution from its i.i.d. samples, which corresponds to the case that W is a constant function.

5.4.1 Agents with Known Grid Positions

In this section, we provide the convergence result of the estimation procedure in Eqn. (5) in the setting where the agents' positions $\{\xi_i\}_{i=1}^N$ form a known grid on $[0,1]$. Without loss of generality, we assume that $\xi_i \le \xi_j$ for any $i \le j$ in $[N]$, and denote the set of positions as $\bar{\xi} = \{\xi_i\}_{i=1}^N$. In this section, our behavior policies $\pi_\tau^{\mathcal{I}}$ for $\tau \in [L]$ are set as L_π-Lipschitz policies. It means that $\|\pi_h^\alpha(\cdot \,|\, s) - \pi_h^\beta(\cdot \,|\, s)\|_1 \le L_\pi|\alpha - \beta|$ for all $h \in [H]$ and $\alpha, \beta \in \mathcal{I}$. We note that setting the behavior policies as Lipschitz policies will not restrict the applicability of our estimation procedure, since the NE is shown to be Lipschitz under Assumptions 1 and 5 in Appendix P.

Then we introduce the performance metric for our estimates. Given $\bar{\xi}$, the joint distribution of $(s_{\tau,h}^i, a_{\tau,h}^i, \mu_{\tau,h}^{\mathcal{I}}, r_{\tau,h}^i, s_{\tau,h+1}^i)_{i=1}^N$ is $\prod_{i=1}^N \rho_{\tau,h}^i$, where $\rho_{\tau,h}^i = \mu_{\tau,h}^i \times \pi_{\tau,h}^i \times \delta_{\mu_{\tau,h}^{\mathcal{I}}} \times \delta_{r_h^*} \times P_h^*$. Here $\delta_{r_h^*}$ is the delta distribution induced by the deterministic function r_h^*. We define the risk of (f, g, W) given $\bar{\xi}$ as

$$\mathcal{R}_{\bar{\xi}}(f, g, W) = \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^i}\Big[\Big(s_{\tau,h+1}^i - f\big(\omega_{\tau,h}^i(W)\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\omega_{\tau,h}^i(W)\big)\Big)^2\Big]. \quad (7)$$

The risk $\mathcal{R}_{\bar{\xi}}(f, g, W)$ measures the mean square error of the estimates f, g, W with respect to the distributions of states, actions and distribution flow on the sampled agents. This risk definition is motivated by the distribution regression (Szabó et al., 2015), since our framework is a generalization of the distribution regression, as discussed in Section 5.4. The convergence rate of our estimates $(\hat{f}_h, \hat{g}_h, \hat{W}_h)$ is stated as follows.

Theorem 5 *Under Assumptions 1, 5, 6, and 7, if $\{\xi_i\}_{i=1}^N$ are known grid positions such that $\xi_i = i/N$ for $i \in [N]$, then with probability at least $1-\delta$, the risk of the estimates in Eqn.* (5) *can be bounded as*

$$\begin{aligned}&\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)\\ &\quad= O\bigg(\underbrace{\frac{(B_S + \bar{r}B_K)^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}}_{\text{generalization error}} + \underbrace{\frac{(B_S + \bar{r}B_K)\bar{r}L_K B_k}{\sqrt{N}}\log\frac{NL\mathcal{N}_\infty(1/\sqrt{N}, \tilde{\mathcal{W}})}{\delta}}_{\text{mean-embedding estimation error}}\bigg),\end{aligned}$$

where

$$N_{\mathbb{B}_r} = \mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{3}{NL}, \mathbb{B}(r, \bar{\mathcal{H}})\bigg), \quad N_{\tilde{\mathbb{B}}_{\tilde{r}}} = \mathcal{N}_{\tilde{\mathcal{H}}}\bigg(\frac{3}{NL}, \mathbb{B}(\tilde{r}, \tilde{\mathcal{H}})\bigg), \quad N_{\tilde{\mathcal{W}}} = \mathcal{N}_\infty\bigg(\frac{3}{L_K NL}, \tilde{\mathcal{W}}\bigg).$$

The proof of the theorem and the definitions of these covering numbers are provided in Appendix G. The estimation error in Theorem 5 consists of two terms: the first term corresponds to the generalization error, and the second term corresponds to the mean-embedding estimation error. The generalization error involves the error from optimizing over the empirical mean of the risk in Eqn. (5) instead of the population risk in Eqn. (7). The mean-embedding estimation error comes from the fact that we cannot directly observe the distribution flow $\mu_\tau^{\mathcal{I}}$, but we need to estimate it from the states of sampled agents. As discussed in Section 5.4, the mean-embedding estimation error consists of the agent sampling error and the state sampling error. If we use finite general function classes, then the covering number in the bound will be replaced by the cardinalities of these function classes. The resultant convergence rate would thus be $O(1/\sqrt{N})$.

The model learning algorithm in Pasztor et al. (2021) for the MFG assumes access to the nominal value of the distribution flow. Such an assumption can be achieved in MFGs by sampling a large number of agents at each time, since all the agents are homogeneous and have the same state distribution flow. This estimation procedure will however, come at a cost of $O(1/\sqrt{N})$, which is not reflected in their results. What's more, such an assumption is no longer realistic in the GMFG, since the agents in GMFG are heterogeneous, and the state distributions of agents are different. Our estimation procedure in (5) does not require the access to the nominal value of the distribution flow $\mu_h^{\mathcal{I}}$. Instead, we estimate this quantity from states of sampled agents and prove that such an estimate works for the heterogeneous agents.

5.4.2 Agents with Known Random Positions

In this section, we provide the convergence result of estimation procedure in Eqn. (5) in the setting where the agent positions $\{\xi_i\}_{i=1}^N$ are known realizations of i.i.d. samples drawn from Unif($[0,1]$). The set of positions is denoted as $\bar{\xi} = \{\xi_i\}_{i=1}^N$. We first specify the performance metric in this section. For an agent $\alpha \in \mathcal{I}$, we denote the joint distribution of $(s_{\tau,h}^\alpha, a_{\tau,h}^\alpha, \mu_{\tau,h}^{\mathcal{I}}, r_{\tau,h}^\alpha, s_{\tau,h+1}^\alpha)$ as $\rho_{\tau,h}^\alpha$, where $\rho_{\tau,h}^\alpha = \mu_{\tau,h}^\alpha \times \pi_{\tau,h}^\alpha \times \delta_{\mu_{\tau,h}^{\mathcal{I}}} \times \delta_{r_h^*} \times P_h^*$. Then the risk of $f \in \mathcal{H}$, $g \in \tilde{\mathcal{H}}$, and $\tilde{W} \in \tilde{\mathcal{W}}$ is defined as

$$\mathcal{R}(f,g,\tilde{W}) = \frac{1}{L}\sum_{\tau=1}^{L}\int_0^1 \mathbb{E}_{\rho_{\tau,h}^\alpha}\biggl[\Bigl(s_{\tau,h+1}^\alpha - f\bigl(\omega_{\tau,h}^\alpha(\tilde{W})\bigr)\Bigr)^2 + \Bigl(r_{\tau,h}^\alpha - g\bigl(\omega_{\tau,h}^\alpha(\tilde{W})\bigr)\Bigr)^2\biggr]\mathrm{d}\alpha. \quad (8)$$

Compared to the risk with grid positions defined in Eqn. (7), the risk defined in Eqn. (8) can be derived by taking expectation with respect the distribution of the positions, i.e., $\mathcal{R}(f,g,W) = \mathbb{E}_{\bar{\xi}}\bigl[\mathcal{R}_{\bar{\xi}}(f,g,W)\bigr]$. The convergence rate of our estimates can be stated as follows.

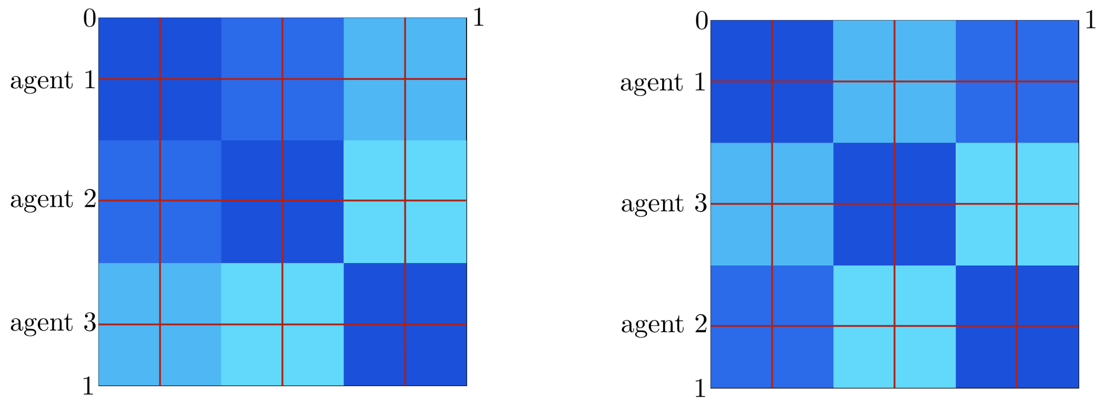
Theorem 6 *Under Assumptions 5, 6, 7, and 1, if $\{\xi_i\}_{i=1}^N$ are known i.i.d. samples of Unif($[0,1]$), then with probability at least $1-\delta$, the risk of the estimates in Eqn.* (5) *can be bounded as*

$$\begin{aligned}
&\mathcal{R}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}(f_h^*, g_h^*, W_h^*) \\
&\quad = O\biggl(\frac{(B_S + \bar{r}B_K)^2}{\sqrt{N}}\log\frac{\tilde{N}_{\mathbb{B}_r}\tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}}\tilde{N}_{\tilde{\mathcal{W}}}}{\delta} \\
&\qquad\quad + \frac{(B_S + \bar{r}B_K)\bar{r}L_K B_k}{\sqrt{N}}\log\frac{NL\mathcal{N}_\infty\bigl(1/\sqrt{N}, \tilde{\mathcal{W}}\bigr)}{\delta} + \frac{(B_S + \bar{r}B_K)^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}\biggr),
\end{aligned}$$

where

$$\tilde{N}_{\mathbb{B}_r} = \mathcal{N}_{\bar{\mathcal{H}}}\biggl(\frac{1}{16\sqrt{N}}, \mathbb{B}(r,\bar{\mathcal{H}})\biggr), \tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}} = \mathcal{N}_{\tilde{\mathcal{H}}}\biggl(\frac{1}{16\sqrt{N}}, \mathbb{B}(\tilde{r},\tilde{\mathcal{H}})\biggr), \tilde{N}_{\tilde{\mathcal{W}}} = \mathcal{N}_\infty\biggl(\frac{1}{16rL_K B_k\sqrt{N}}, \tilde{\mathcal{W}}\biggr).$$

The proof is provided in Appendix H. The estimation error in Theorem 6 consists of three terms: the first term corresponds to the approximation error, the second term corresponds to the generalization error, and the third term corresponds to the mean-embedding estimation error. The first term comes from the fact that we can only approximate the risk $\mathcal{R}(f,g,W)$ by $\mathcal{R}_{\bar{\xi}}(f,g,W)$ in the estimation procedure specified in Eqn. (5). The second and the third terms can be explained in the same way as for the terms in Theorem 5.



(a) The SBM graphon and three sampled agents.

(b) Transformed SBM graphon and the correspondingly transformed agents.

Figure 1: The left figure shows the SBM graphon and three sampled agents. Swapping the second and the third communities, we obtain the graphon on the right. The sampled agents are correspondingly swapped. Although the graphons and agent positions in the left and the right figures are not the same, the agents in both figures retain the same "relative positions" with the underlying graphons.

5.5 Learning from Sampled Agents with Unknown Positions

We now consider the setting where the positions of the sampled agents $\{\xi_i\}_{i=1}^N$ are on the grid in $[0,1]$, but are *unknown*. This means that the set of sampled positions $\{\xi_i\}_{i=1}^N$ is equal to $\{i/N\}_{i=1}^N$, but we do not know which i/N each ξ_i corresponds to. In addition to the data collection procedures in Section 5.1, we assume that we implement the *same* policy over L independent rounds. This sampling method implies that the distribution defined in Section 5.5 satisfies $\rho_{\tau,h}^{\alpha} = \rho_{\tau',h}^{i}$ for all $\tau, \tau' \in [L]$, $\alpha \in \mathcal{I}$ and $h \in [H]$.

Intuitively, since the position information is missing from our observations, we cannot estimate the precise values of graphons. For example, the collected data from the agents in Figure 1(a) is same as that in Figure 1(b), so we cannot distinguish between these two different graphons. However, we can see that these two graphons are the same up to a measure-preserving bijection. Proposition 2 shows that the model with transformed graphons is the same as the original model up to a measure-preserving bijection. Thus, in this section, our goal is to estimate the model of GMFG up to a measure-preserving bijection.

In this setting, we cannot estimate the mean-embedding $\omega_{\tau,h}^{i}(W_h^*)$ as Eqn. (4), since we do not know the agents' positions $\{\xi_i\}_{i=1}^N$. Instead, we need to estimate the "relative positions" of these agents. Here the relative positions refer to the relationship between the agents' positions and the underlying graphon. For example, in Figure 1, the agents retain the same relative positions in different graphons. With N sampled agents, the relative positions can be represented by the permutation of these agents. We denote the set of all the permutations of N objects as $\mathcal{C}^N$, where $|\mathcal{C}^N| = N!$. For a permutation $\sigma \in \mathcal{C}^N$ and a graphon W, we estimate the relative position of i^{th} agent as $\sigma(i)/N$ for all $i \in [N]$. Then mean-field embedding estimate can be derived as

$$\hat{\hat{\omega}}^{i,\sigma}_{\tau,h}(W) = \frac{1}{(N-1)L}\sum_{j\neq i}\sum_{\tau'=1}^{L} W\left(\frac{\sigma(i)}{N}, \frac{\sigma(j)}{N}\right) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s^j_{\tau',h})\big). \tag{9}$$

Similar as Eqn. (14), Eqn. (9) is also an average over L episodes, since we implement the same policy for L independent times. In this estimate, only the relative positions between agents and the underlying graphon matters, so we can equivalently express such estimate with a transformed graphon. We define $\hat{\hat{\omega}}^{i}_{\tau,h}(W)$ as $\hat{\hat{\omega}}^{i,\sigma}_{\tau,h}(W)$ with the identity map σ. The set of measure-preserving bijections that are permutations of the intervals $[(i-1)/N, i/N]$ for $i \in [N]$ is denoted as $\mathcal{C}^N_{[0,1]}$. Then for some $\phi \in \mathcal{C}^N_{[0,1]}$, the estimate in Eqn. (9) can be reformulated as

$$\hat{\hat{\omega}}^{i}_{\tau,h}(W^\phi) = \frac{1}{(N-1)L}\sum_{j\neq i}\sum_{\tau'=1}^{L} W\Big(\phi\big(i/N\big), \phi\big(j/N\big)\Big) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s^j_{\tau',h})\big).$$

Given this mean-embedding estimate, our model estimation estimation procedure can be stated as

$$(\hat{f}_h, \hat{g}_h, \hat{W}_h, \hat{\phi}_h) = \mathop{\mathrm{argmin}}_{\substack{f\in\mathbb{B}(r,\bar{\mathcal{H}}),\\ g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}),\\ \tilde{W}\in\tilde{\mathcal{W}},\phi\in\mathcal{C}^N_{[0,1]}}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - f\big(\hat{\hat{\omega}}^i_{\tau,h}(\tilde{W}^\phi)\big)\Big)^2 + \Big(r^i_{\tau,h} - g\big(\hat{\hat{\omega}}^i_{\tau,h}(\tilde{W}^\phi)\big)\Big)^2. \tag{10}$$

We note that the computational burden of this procedure can be high due the large number of permutations in $\mathcal{C}^N_{[0,1]}$ (for large N). However, this high computational burden is common in the graphon learning algorithms (Gao et al., 2015; Klopp et al., 2017). Our work mainly focuses on the statistical analysis of the graphon problem, rather than their well-known computational limitations. We leave the addressing of computational concerns to future work. We then specify the performance metric under this setting. As mentioned earlier, we cannot estimate the precise values of graphons. Thus, we measure the accuracy of our estimates by transforming the graphon estimate with the optimal measure-preserving bijections. Such a risk is known as the *permutation-invariant risk*, which is defined as

$$\begin{aligned}\bar{\mathcal{R}}_{\bar{\xi}}(f,g,W) &= \inf_{\phi\in\mathcal{B}_{[0,1]}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\bigg[\Big(s^i_{\tau,h+1} - f\big(\omega^i_{\tau,h}(W^\phi)\big)\Big)^2 + \Big(r^i_{\tau,h} - g\big(\omega^i_{\tau,h}(W^\phi)\big)\Big)^2\bigg]\\ &= \inf_{\phi\in\mathcal{B}_{[0,1]}} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_h}\bigg[\Big(s^i_{h+1} - f\big(\omega^i_{h}(W^\phi)\big)\Big)^2 + \Big(r^i_{\tau,h} - g\big(\omega^i_{\tau,h}(W^\phi)\big)\Big)^2\bigg],\end{aligned} \tag{11}$$

where $\rho^i_h = \rho^i_{\tau,h}$ for all $\tau \in [L]$. The term “permutation-invariant” comes from the analogy between permutations and measure-preserving bijections and the fact that $\bar{\mathcal{R}}_{\bar{\xi}}(f,g,W) = \bar{\mathcal{R}}_{\bar{\xi}}(f,g,W^\phi)$ for any $\phi \in \mathcal{B}_{[0,1]}$. In the graphon learning problem, similar distances that take the permutation-invariance into account have been defined in existing works (Klopp and Verzelen, 2019). Our risk here is designed to reflect the permutation-invariance property and various GMFG-related quantities. Our convergence guarantee of the estimation procedure can be stated as follows.

Theorem 7 *Under Assumptions 5, 6, 7, and 1, if $\{\xi_i\}_{i=1}^N = \{i/N\}_{i=1}^N$, then with probability at least $1-\delta$, the risk of the estimates in Eqn.* (10) *can be bounded as*

$$\begin{aligned}&\bar{\mathcal{R}}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \bar{\mathcal{R}}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)\\ &\quad = O\bigg(\underbrace{\frac{L_{\mathcal{W}} B_k \bar{r} L_K (B_S + \bar{r}B_K)}{N}}_{\text{agent sampling error}} + \underbrace{(B_S + \bar{r}B_K)\bar{r}L_K B_K\sqrt{\frac{N}{L}}\log\frac{NL\mathcal{N}_\infty(\sqrt{N/L}, \tilde{\mathcal{W}})}{\delta}}_{\text{state sampling error}}\\ &\qquad + \underbrace{\frac{(B_S + \bar{r}B_K)^4}{L}\log\frac{N\tilde{N}_{\mathbb{B}_r}\tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}}\tilde{N}_\infty}{\delta}}_{\text{generalization error}}\bigg),\end{aligned}$$

where

$$\tilde{N}_{\mathbb{B}_r} = \mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{3}{L}, \mathbb{B}(r, \bar{\mathcal{H}})\bigg), \quad \tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}} = \mathcal{N}_{\tilde{\mathcal{H}}}\bigg(\frac{3}{L}, \mathbb{B}(\tilde{r}, \tilde{\mathcal{H}})\bigg), \quad \tilde{N}_{\tilde{\mathcal{W}}} = \mathcal{N}_\infty\bigg(\frac{3}{L_K L}, \tilde{\mathcal{W}}\bigg).$$

The proof is provided in Appendix I. The estimation error in Theorem 7 consists of three terms: the first two terms correspond to the mean-embedding estimation error, and the last term corresponds to the generalization error. As mentioned in Section 5.4, the mean-embedding estimation error consists of agent sampling error and the state sampling error. The first term in the bound represents the agent sampling error. Since the distance between adjacent agents is $1/N$, this approximation error is of order $O(1/N)$. The second term represents the state sampling error. The term $\sqrt{N}$ in the numerator comes from the estimation of relative positions from $\mathcal{C}^N$, whose size is $N!$, and the union bound among this set. The third term, which is the generalization error, also suffers from the union bound of $N!$ relative positions. Compared with Corollary 12 in Section 5.4, the result in Theorem 7 suffers from a multiplicative factor $\log N!$. When the function classes are finite and $L = \Theta(N^\beta)$ with $\beta > 1$, the convergence rate in Theorem 7 is $O(\max\{N^{-(\beta-1)/2}, N^{-1}\})$. In contrast, the convergence rate Corollary 12 is $O(N^{-(\beta+1)/2})$.

Theorem 7 states the estimate error in the permutation-invariant risk. In fact, we can also derive the convergence rate of our estimation of relative positions $\hat{\phi}_h$. This means that for some unknown correction $\psi^* \in \mathcal{C}_{[0,1]}^N$, the risk defined in Eqn. (7) of our estimate $(\hat{f}_h, \hat{g}_h, \hat{W}_h^{\hat{\phi}_h \circ \psi^*})$ vanishes.

Corollary 8 *Given $\{\xi_i\}_{i=1}^N = \{i/N\}_{i=1}^N$, we adopt $\psi^* \in \mathcal{C}_{[0,1]}^N$ to denote the mapping that $\psi^*(\xi_i) = i/N$ for all $i \in [N]$. Under Assumptions 5, 6, 7, and 1, the risk of estimate can be bounded as*

$$\begin{aligned}&\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h^{\hat{\phi}_h \circ \psi^*}) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)\\ &\quad = O\bigg(\frac{L_{\mathcal{W}} B_k \bar{r} L_K (B_S + \bar{r}B_K)}{N} + (B_S + \bar{r}B_K)\bar{r}L_K B_K\sqrt{\frac{N}{L}}\log\frac{NL\mathcal{N}_\infty(\sqrt{N/L}, \tilde{\mathcal{W}})}{\delta}\\ &\qquad + \frac{(B_S + \bar{r}B_K)^4}{L}\log\frac{N\tilde{N}_{\mathbb{B}_r}\tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}}\tilde{N}_\infty}{\delta}\bigg)\end{aligned}$$

with probability at least $1-\delta$.

Algorithm 2 Estimation of $\hat{\mu}_t^{\mathcal{I}}$, $\hat{\bar{\mu}}_{t+1}^{\mathcal{I}}$, and $\hat{Q}_h^{\lambda,\alpha}(s,a,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})$

Inputs: the current policy $\pi_t^{\mathcal{I}}$ and the past distribution flow estimate $\hat{\bar{\mu}}_t^{\mathcal{I}}$
Outputs: $\hat{\mu}_t^{\mathcal{I}}$, $\hat{\bar{\mu}}_{t+1}^{\mathcal{I}}$, and $\hat{Q}_h^{\lambda,\alpha}(\cdot,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})$ for all $h\in[H]$, $\alpha\in\mathcal{I}$
Procedure:

1: Implement policy $\pi_t^{\mathcal{I}}$ for L times and collect the data $\{\mathcal{D}_\tau\}_{\tau=1}^L$ (with any kind of sampled agents in Section 5)
2: Derive the MDP estimate $(\hat{P},\hat{r},\hat{W})$ with the estimation procedures in Section 5, where $\hat{P}=\{\hat{P}_h\}_{h=1}^H$, $\hat{P}=\{\hat{P}_h\}_{h=1}^H$, $\hat{r}=\{\hat{r}_h\}_{h=1}^H$, and $\hat{W}=\{\hat{W}_h\}_{h=1}^H$
3: Derive $\hat{\mu}_t^{\mathcal{I}}$ as the distribution flow of implementing $\pi_t^{\mathcal{I}}$ on the MDP estimate.
4: Derive $\hat{\bar{\mu}}_{t+1}^{\mathcal{I}}$ as $\hat{\bar{\mu}}_{t+1}^{\mathcal{I}}=(1-\alpha_t)\hat{\bar{\mu}}_t^{\mathcal{I}}+\alpha_t\hat{\mu}_t^{\mathcal{I}}$.
5: Implement a behavior policy $\pi_t^{\mathrm{b},\mathcal{I}}$ on the MDP induced by $\hat{\bar{\mu}}_t^{\mathcal{I}}$ for L times and collect the data $\{\mathcal{D}'_\tau\}_{\tau=1}^L$ (with any kind of sampled agents in Section 5)
6: Derive the MDP estimate $(\hat{P}',\hat{r}',\hat{W}')$ with the estimation procedures in Section 5
7: Derive $\hat{Q}_h^{\lambda,\alpha}(\cdot,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})$ as action-value functions of $\pi_t^{\mathcal{I}}$ on the MDP estimate $(\hat{P}',\hat{r}',\hat{W}')$.

The proof is provided in Appendix M. Combined with Proposition 2, Corollary 8 shows that the model estimate $(\hat{f}_h,\hat{g}_h,\hat{W}_h^{\hat{\phi}_h})$ converges to the nominal model in the sense that they are shown to be equivalent up to an unknown measure-preserving bijection $\psi^*\in\mathcal{C}_{[0,1]}^N$.

6. Combination of Optimization and Estimation Results

In this section, we make use of the estimator we constructed and analyzed in Section 5 to derive estimates in Step 4 in Algorithm 1. We assume that one has access to a population simulator; this assumption is commonly made in the MFG literature (Guo et al., 2019; Anahtarci et al., 2019, 2022). This simulator is able to generate data according to two types of requests: (i) implement policy $\pi^{\mathcal{I}}$ on the MDP induced by a pre-specified distribution flow $\mu^{\mathcal{I}}$, (ii) implement policy $\pi^{\mathcal{I}}$ directly. In the latter case, the MDP is induced by the distribution flow of $\pi^{\mathcal{I}}$ itself.

We use Algorithm 2 to derive the distribution flow and action-value function estimate in Line 4 of Algorithm 1. In this algorithm, we call the simulator twice. First, we directly implement the policy $\pi_t^{\mathcal{I}}$ for L times independently. With the collected data, we can estimate the distribution flow $\mu_t^{\mathcal{I}}$. Second, we implement a behavior policy $\pi_t^{\mathrm{b},\mathcal{I}}$ on the MDP induced by $\hat{\bar{\mu}}_t^{\mathcal{I}}$ for L times. Then estimate the action-value functions with the collected data.

One natural question is that why we need to estimate the transition kernels and underlying graphons to estimate $\mu_t^{\mathcal{I}}$. An alternative is to implement $\pi_t^{\mathcal{I}}$ for L times and estimate the distribution flow of the sampled agents as their empirical distribution. In fact, the convergence rate of the alternative will be $O(1/\sqrt{L})$ from central limit theorem. However, our estimate will shown to have risk bounded by $O(1/\sqrt{NL})$. This improvement is because our algorithm makes use of the information of all the agents, but the alternative only uses the information of single agent for the estimation.

To derive the theoretical guarantees on the accuracy of the distribution flow and action-value function estimates, we make the following assumptions.

Assumption 8 *There exist $L_\varepsilon > 0$ such that the noises ε_h for $h \in [H]$ satisfy that for any $a \in \mathbb{R}$, $\mathrm{TV}(\varepsilon_h + a, \varepsilon_h) \leq L_\varepsilon a$ for all $h \in [H]$.*

This assumption enables us to control the total variation error of our transition kernels P_h^* by the estimation error of f_h^*. We note that Assumption 8 is satisfied for a wide range of distributions, including the uniform distribution, the centralized Beta distributions for $\alpha > 1, \beta > 1$, and the truncated Gaussian distribution. We then assume that the behavior policy $\pi_t^{\mathrm{b},\mathcal{I}}$ satisfies the following assumptions.

Assumption 9 *There exist two constants $C_\pi, C'_\pi > 0$ such that for all $t \in [T]$*

$$\sup_{s\in\mathcal{S},a\in\mathcal{A},\alpha\in\mathcal{I},h\in[H]} \frac{\bar{\pi}_{t,h}^{*,\alpha}(a \mid s)}{\pi_{t,h}^{\mathrm{b},\alpha}(a \mid s)} \leq C_\pi \quad \textit{and} \quad \sup_{s\in\mathcal{S},a\in\mathcal{A},\alpha\in\mathcal{I},h\in[H]} \frac{\pi_{t+1,h}^{\alpha}(a \mid s)}{\pi_{t,h}^{\mathrm{b},\alpha}(a \mid s)} \leq C'_\pi.$$

This assumption states that the behavior policy should explore the actions of the NE and the policy $\pi_{t+1}^{\mathcal{I}}$. It is quite natural since we want to estimate the action-value function of $\pi_{t+1}^{\mathcal{I}}$ from the data collected by $\pi_t^{\mathrm{b},\mathcal{I}}$. Similar assumptions have been commonly made in the off-policy evaluation literature (Kallus et al., 2021; Uehara et al., 2020).

Assumption 10 *For any policy $\pi^{\mathcal{I}} \in \tilde{\Pi}$, we define $\mu^{+,\mathcal{I}} = \Gamma_3(\pi^{\mathcal{I}}, \bar{\mu}_t^{\mathcal{I}}, W^*)$. We also define $\bar{\mu}_t^{\mathrm{b},\mathcal{I}} = \Gamma_3(\pi_t^{\mathrm{b},\mathcal{I}}, \bar{\mu}_t^{\mathcal{I}}, W^*)$. There exists a constant $C''_\pi > 0$ such that for any $t \in [T]$ and any policy $\pi^{\mathcal{I}}$ specified above, we have*

$$\sup_{s\in\mathcal{S},h\in[H],\alpha\in\mathcal{I}} \frac{\mu_h^{+,\alpha}(s)}{\bar{\mu}_{t,h}^{\mathrm{b},\alpha}(s)} \leq C''_\pi.$$

This assumption states that the behavior policy should be sufficiently exploratory such that the induced distribution of other policies can be covered by that of the behavior policy. Similar assumptions haven been made in the policy optimization literatures (Shani et al., 2020; Agarwal et al., 2020). We note that if we take the behavior policy $\pi_t^{\mathrm{b},\mathcal{I}} = \mathrm{Unif}(\mathcal{A})^{\mathcal{I}\times H}$ to be the uniform distribution on the action space, then the constants in Assumptions 9 and 10 can be set as $C_\pi = C'_\pi = |\mathcal{A}|$ and $C''_\pi = |\mathcal{A}|^H$.

6.1 Known-position Case

In this section, we analyze Algorithm 1 and Algorithm 2 when we know the positions (grid or random) of the sampled agents. In Algorithm 2, we know the distribution flow $\hat{\bar{\mu}}_t^{\mathcal{I}}$ during our second call of the simulator. Thus, in Line 6 of Algorithm 2, we estimate the model from the collected and the precise value of the distribution flows. This estimation procedure can be acquired by simplifying the estimation procedure in Section 5.4.1 as

$$(\hat{f}_h, \hat{g}_h, \hat{W}_h) = \operatorname*{argmin}_{\substack{f\in\mathbb{B}(r,\tilde{\mathcal{H}}),\\ g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}),\\ \tilde{W}\in\tilde{\mathcal{W}}}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s_{\tau,h+1}^i - f\big(\omega_{\tau,h}^i(\tilde{W})\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\omega_{\tau,h}^i(\tilde{W})\big)\Big)^2, \quad (12)$$

where $\omega_{\tau,h}^i(W)$ is the mean-embedding calculated by Eqn. (3) and the known distribution flow. Then the result for the agents with known grid positions is stated as

Corollary 9 *If we sample agents with known grid positions and adopt Algorithms* (5) *and* (12) *to estimate the MDP, then under Assumptions 1 to 10, we have that GMFG-PPO return the following estimates with probability at least* $1-\delta$

$$D\bigg(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\mathcal{I}},\pi^{*,\mathcal{I}}\bigg)+d\bigg(\frac{1}{T}\sum_{t=1}^{T}\hat{\mu}_t^{\mathcal{I}},\mu^{*,\mathcal{I}}\bigg)$$
$$=O\bigg(\frac{B_S+\bar{r}B_K}{(NL)^{\frac{1}{4}}}\log^{\frac{1}{4}}\frac{TN_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}+\frac{(B_S+\bar{r}B_K)^{\frac{1}{4}}(\bar{r}L_KB_k)^{\frac{1}{4}}}{(NL)^{\frac{1}{8}}}\log^{\frac{1}{4}}\frac{TNL\mathcal{N}_\infty(1/\sqrt{N},\tilde{\mathcal{W}})}{\delta}$$
$$+\frac{(B_S+\bar{r}B_K)^{\frac{1}{4}}(B_S+\bar{r}B_K+\bar{r}L_KB_k)^{\frac{1}{4}}}{N^{\frac{1}{4}}}\bigg)+O\bigg(\frac{\sqrt{\log T}}{T^{1/3}}\bigg).$$

The proof is provided in Appendix K. The error of learning NE consists of two types of terms. The first originates from the estimation error of the distribution flow and the action-value function. It involves the number of sampled agents N and the number of episodes L. The second represents the optimization error and involves the number of iterations T. Consider the case where the function classes are finite. To learn a NE with error ε measured according to $D(\cdot,\cdot)$ and $d(\cdot,\cdot)$, we can run Algorithms 1 and 2 with $T=\tilde{O}(\varepsilon^{-3})$ iterations and $O((NL)^{-1/8}+N^{-1/4})=\varepsilon$. The second condition can be achieved by several parameter settings, e.g., $L=1$, $N=O(\varepsilon^{-8})$ and $L=O(\varepsilon^{-4})$, $N=O(\varepsilon^{-4})$.

The result for the agents with known random positions is stated as follows.

Corollary 10 *If we sample agents with known random positions and adopt Algorithms* (5) *and* (12) *to estimate the MDP, then under Assumptions 1 to 10, we have that GMFG-PPO return the following estimates with probability at least* $1-\delta$

$$D\bigg(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\mathcal{I}},\pi^{*,\mathcal{I}}\bigg)+d\bigg(\frac{1}{T}\sum_{t=1}^{T}\hat{\mu}_t^{\mathcal{I}},\mu^{*,\mathcal{I}}\bigg)$$
$$=O\bigg(\frac{B_S+\bar{r}B_K}{(NL)^{\frac{1}{4}}}\log^{\frac{1}{4}}\frac{TN_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}+\frac{(B_S+\bar{r}B_K)^{\frac{1}{4}}(\bar{r}L_KB_k)^{\frac{1}{4}}}{(NL)^{\frac{1}{8}}}\log^{\frac{1}{4}}\frac{TNL\mathcal{N}_\infty(1/\sqrt{NL},\tilde{\mathcal{W}})}{\delta}$$
$$+\frac{(B_S+\bar{r}B_K)^{1/2}}{N^{\frac{1}{8}}}\log^{\frac{1}{4}}\frac{\tilde{N}_{\mathbb{B}_r}\tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}}\tilde{N}_{\tilde{\mathcal{W}}}}{\delta}\bigg)+O\bigg(\frac{\sqrt{\log T}}{T^{1/3}}\bigg).$$

The proof is provided in Appendix L. Similar as Corollary 9, error of learning NE consists of the estimation error and the optimization error. To learn a NE with error ε measured according to $D(\cdot,\cdot)$ and $d(\cdot,\cdot)$, we can run Algorithms 1 and 2 with $T=\tilde{O}(\varepsilon^{-3})$ iterations and $N=O(\varepsilon^{-8})$ sampled agents.

6.2 Unknown-position Case

In this section, we analyze Algorithms 1 and 2 when we do not know the grid positions of the sampled agents. In Algorithm 2, we need to specify the policy $\pi_t^{\mathcal{I}}$ and distribution flow $\hat{\mu}_t^{\mathcal{I}}$, which requires the information of agents' positions. Thus, we additionally assume that for a specific agent $\alpha\in\mathcal{I}$, we know which sampled agent is closest to α and the relative position to the closest sampled agent. This assumption holds in many realistic scenarios.

For example, we consider the swarm robotics related problems (Elamvazhuthi and Berman, 2019). In this problem, we would like to find the NE of swarm robotics. The state and action are the kinetic signals and acceleration of robotics, respectively. The reward is the quantity related to the kinetic goal. The robotics that have close physical positions usually share close positions in the underlying graphon, since the interaction among the swarm robotics is related to the physics setting of them. Thus, in this example, although we do not know their exact positions in graphons, we have information about their relative closeness in graphons via their physical positions. In addition, since the data points are stored in each robotics, the samples across different iterations can be guaranteed to come from the same robotics. In this case, there is one sampled person from each state, and we assume that each person knows which state she belongs to, i.e., which sampled person is the closest person to her.

Corollary 11 *If we sample agents with known grid positions and adopt Algorithms* (5) *and* (12) *to estimate the MDP, then under Assumptions 1 to 10, we have that GMFG-PPO return the following estimates with probability at least* $1-\delta$

$$
\begin{aligned}
&D\bigg(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\mathcal{I}},\pi^{*,\mathcal{I}}\bigg)+d\bigg(\frac{1}{T}\sum_{t=1}^{T}\hat{\bar{\mu}}_t^{\mathcal{I}},\mu^{*,\mathcal{I}}\bigg)\\
&\quad=O\bigg(\frac{B_k\bar{r}L_K(B_S+\bar{r}B_K)}{N^{1/4}}+\frac{(B_S+\bar{r}B_K)^{1/4}(\bar{r}L_KB_KN)^{1/4}N^{1/8}}{L^{1/8}}\log^{1/4}\frac{NL\mathcal{N}_\infty(\sqrt{N/L},\tilde{\mathcal{W}})}{\delta}\\
&\qquad+\frac{B_S+\bar{r}B_K}{L^{1/4}}\log^{1/4}\frac{N\tilde{N}_{\mathbb{B}_r}\tilde{N}_{\tilde{\mathbb{B}}_{\bar{r}}}\tilde{N}_\infty}{\delta}\bigg)+O\bigg(\frac{\sqrt{\log T}}{T^{1/3}}\bigg).
\end{aligned}
$$

The proof is provided in Appendix N. Similar to Corollaries 9 and 10, the learning error in Corollary 11 consists of the estimation error and the optimization error. To learn a NE with error ε measured according to $D(\cdot,\cdot)$ and $d(\cdot,\cdot)$, we can run Algorithms 1 and 2 with $T=\tilde{O}(\varepsilon^{-3})$ iterations, $N=O(\varepsilon^{-4})$ sampled agents, and $L=O(\varepsilon^{-12})$ episodes.

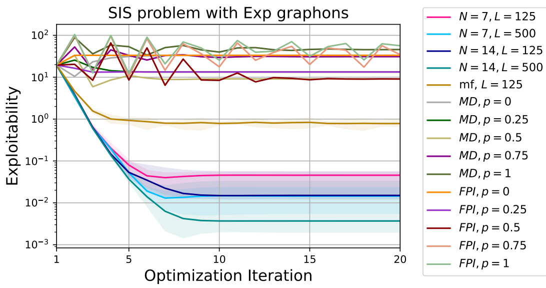
7. Experiments

In this section, we utilize simulations to demonstrate the importance of learning the underlying graphons, thus corroborating our theoretical results. We simulate our algorithms on the Susceptible-Infectious-Susceptible (SIS) problem and investment problem. The detailed definitions of the problems are provided in Appendix A.

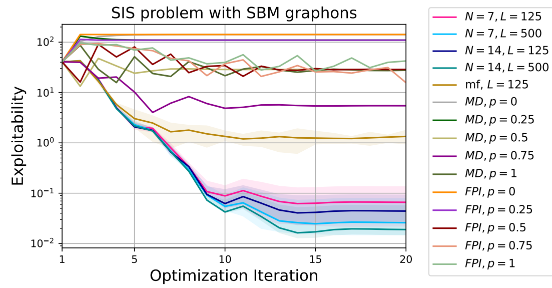
The SIS problem: This problem, which has also been considered in Cui and Koeppl (2021a,b), models the propagation of an epidemic among a large population. People in the population are infected with probability proportional to the number of infected neighbors.

An investment problem: This problem considers the situation where several companies aim to maximize their profits simultaneously. The profit of each company is proportional to the quality of its product and decreases with the total quality of the products in its neighborhood.

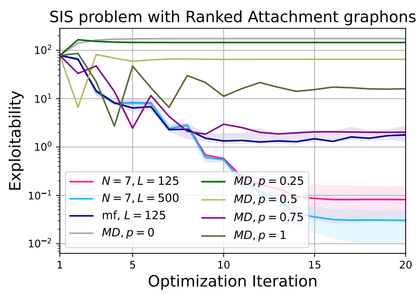
We experiment with four types of graphons: exp-graphon, SBM graphon, affine attachment graphon, and ranked attachment graphon. The value of exp-graphon is affine in the



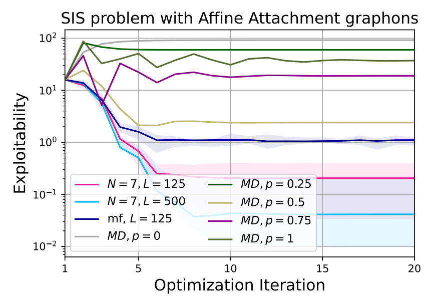
(a) SIS problem with exp-graphons



(b) SIS problem with SBM graphons.



(c) SIS problem with ranked attachment graphons



(d) SIS problem with affine attachment graphons.

Figure 2: Simulation results for SIS problem with grid known-position agents.

exponential of the product of α, β, which is defined as

$$W_\theta^{\exp}(\alpha, \beta) = \frac{2\exp(\theta \cdot \alpha\beta)}{1 + \exp(\theta \cdot \alpha\beta)} - 1, \tag{13}$$

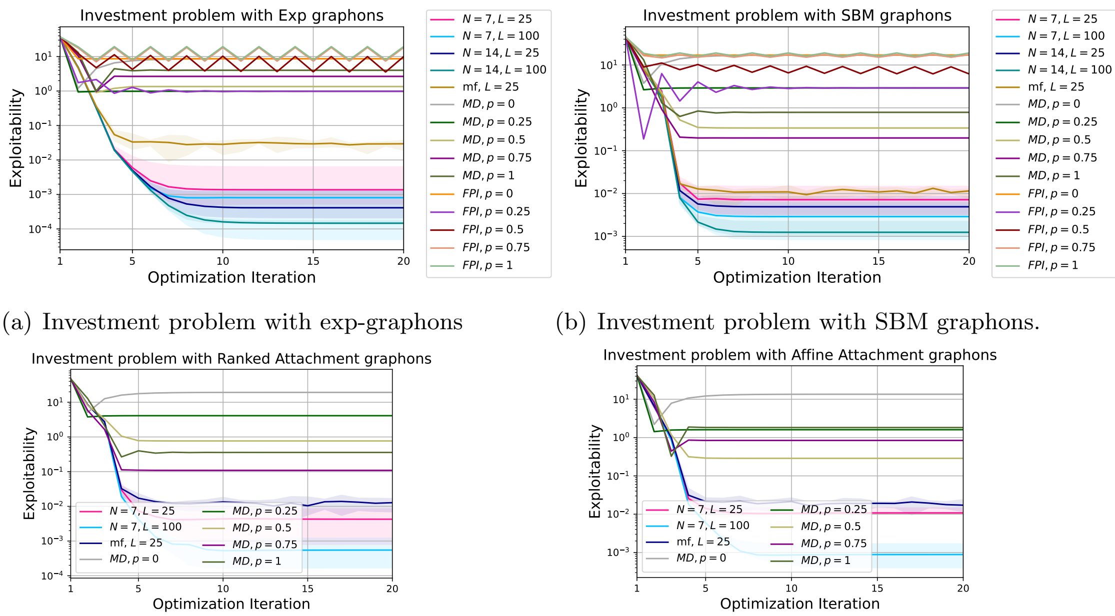
which is parameterized by $\theta > 0$. The SBM graphon splits $[0, 1]$ into $K \geq 1$ blocks, which is parameterized by $\{p_k\}_{k=0}^K$. Here $p_0 = 0$ and $p_K = 1$, and the i-th block is $(p_{i-1}, p_i]$. The value of SBM graphon is then specified by $\{a_{ij}\}_{i,j=1}^{K,K}$ with $a_{ij} = a_{ji}$ as $W^{\text{SBM}}(\alpha, \beta) = a_{ij}$ if $p_{i-1} < \alpha \leq p_i$ and $p_{j-1} < \beta \leq p_j$. The affine attachment graphon is defined as $W_{a,b}^{\text{aff}}(\alpha, \beta) = a - b \cdot (\alpha + \beta)$, where $a, b \in \mathbb{R}$ parameterize the graphon. The ranked attachment graphon is defined as $W_{a,b}^{\text{rank}}(\alpha, \beta) = a - b \cdot \alpha\beta$. This is a generalization of the definition in Cui and Koeppl (2021b).

Since we do not know the nominal value of the NE, we adopt the notion of *exploitability* to measure the closeness between a policy and the NE. For a policy $\pi^{\mathcal{I}}$ and its induced distribution flow $\mu^{\mathcal{I}}$, the exploitability is defined as (Fabian et al., 2022)

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

If we do not learn the underlying graphons, reasonable guesses for them would be constant graphons $W(\alpha, \beta) = p$ for all $\alpha, \beta \in \mathcal{I}$, corresponding to the MFG. In the simulations, we choose the constant p to be $0, 0.25, 0.5, 0.75$ and 1. These values model the cases from the independent agents to the most intensely interacting agents.

Figure 2 displays the exploitability for the algorithms in the SIS problem with different graphons. To learn the system model, we sample $N = 7$ and $N = 14$ agents with known



(a) Investment problem with exp-graphons

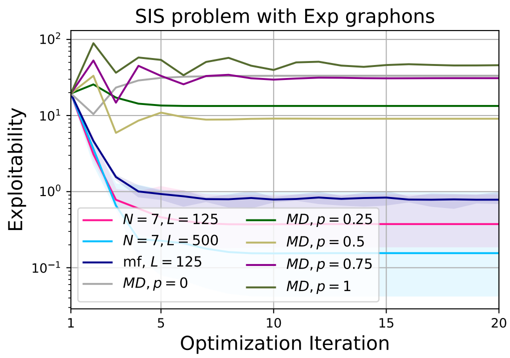
(b) Investment problem with SBM graphons.

(c) Investment problem with ranked attachment graphons

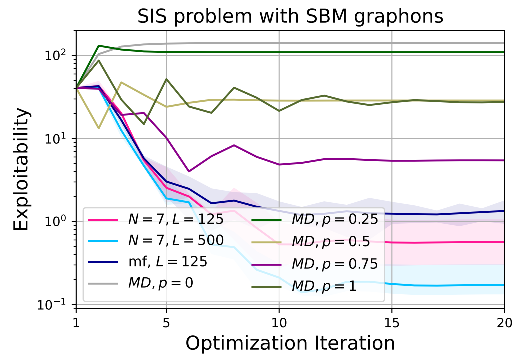
(d) Investment problem with affine attachment graphons.

Figure 3: Simulation results for investment problem with grid known-position agents.

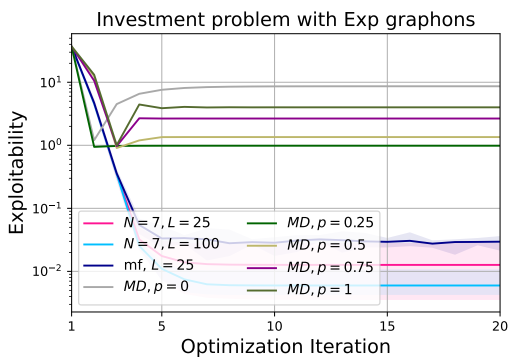
grid positions. The number of episodes for data collection L is set to 125 and 500. The line "mf, $L = 125$" refers to the model-free algorithm in Cui and Koeppl (2021b) that uses 125 trajectories from 7 agents for distribution flow together with value function estimation in each round. "$L = 125$" and "$L = 500$" refer to our algorithms that use 125 and 500 samples from each agent in each round for estimation of the graphons. Since the messages from the results of different graphons are similar, we only display the results for exp graphon and SBM graphon for brevity. Figure 2 demonstrates that our model-based algorithm achieves lower exploitability than the model-free algorithm. The reason is that the estimation error of the model-based algorithm is smaller, as mentioned in Section 6. Lines "$MD, p = 0, 0.25, 0.5, 0.75, 1$" refer to the MFG learning algorithm that implements one-step mirror descent in each iteration Xie et al. (2021); Yardim et al. (2022). Lines "$FPI, p = 0, 0.25, 0.5, 0.75, 1$" refer to the MFG learning algorithm that learns the optimal policy for the current mean-field in each iteration Guo et al. (2019). For these MFG learning algorithms, the reward functions and transition kernels are known to the algorithm. Thus, there are no error bars for these lines. Figure 2 shows that when we assume that the heterogeneous agents are homogeneous, the learning algorithm for NE will suffer from a large error (large exploitability). In contrast, learning the graphons will enable us to learn the NE more accurately. These results demonstrate the necessity of our model learning algorithm in Algorithm 2. We can also observe that the learning error for $N = 7, L = 500$ is less than that for $N = 7, L = 125$, which justifies that the learning error decreases with the increasing trajectory numbers L. In addition, the learning error for $N = 14, L = 125$ is less than that for $N = 7, L = 125$. This shows that the learning error decreases with an increasing number of sampled agents N. These observations corroborate Corollary 9.



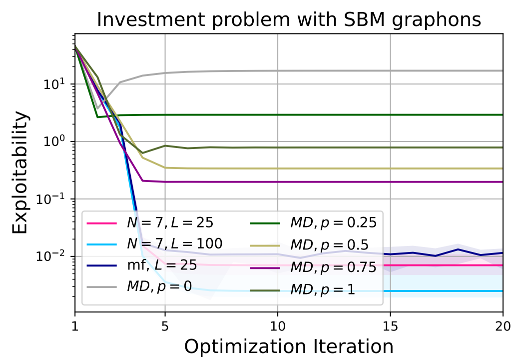
(a) SIS problem with exp-graphons



(b) SIS problem with SBM graphons.



(c) Investment problem with exp-graphons



(d) Investment problem with SBM graphons

Figure 4: Simulation results for SIS and investment problems with random known-position agents.

Figure 3 displays the exploitability for algorithms in the investment problem of Cui and Koeppl (2021b) with different graphons. "$L = 25$" and "$L = 100$" refer to our algorithms that estimate with 125 and 500 samples from each agent in each round. Although the investment problem has a larger state space than the SIS problem, the simulation results contain similar insights as discussed above. These results corroborate Corollary 9.

Figure 4 displays the exploitability for algorithms in the SIS and investment problems with different graphons, where sampled agents have known random positions. We note that lines "$MD, p = 0, 0.25, 0.5, 0.75, 1$" are same as those in Figures 2 and 3, since these MFG algorithms have the full information of the reward function and transition kernel where different estimation setting will not affect the MFG algorithm performance. Figure 4 shows that the GMFG learning algorithms have better performance than the MFG learning algorithm. The reason is that the MFG learning algorithms wrongly assume that all the agents are homogeneous. Figure 4 also indicates that "$N = 7, L = 500$" (resp. "$N = 7, L = 100$") has better performance than "$N = 7, L = 125$" (resp. "$N = 7, L = 25$") in SIS problem (resp. investment problem), which corroborates with Corollary 10.

8. Conclusion

In this paper, we investigated learning the NE of GMFG in the graphons incognizant case. Provably efficient optimization algorithms were designed and analyzed with an estimation oracle, which improved on the previous works in convergence rate. In addition, adopting the mean-embedding ideas, we designed and analyzed the model-based estimation algorithms with sampled agents. Here, the sampled agents have known or unknown positions. These estimation algorithms feature as the first model-based algorithms in GMFG without the distribution flow information. We leave the analysis of more complex agent sampling schemes for future works.

Acknowledgments

This research work is funded by the Singapore Ministry of Education AcRF Tier 2 grant (A-8000423-00-00) and Tier 1 grants (A-8000189-01-00 and A-8000980-00-00).

Appendices for "Learning Graphon Mean-Field Games with Unknown Graphons"

Appendix A. Experimental Details

In this section, we provide the details of our experiments shown in Section 7 of the main paper. We first formally define the SIS and investment problems.

SIS problem: The state space of this problem $\mathcal{S} = \{S, I\}$ consists of the states S (susceptible) and I (infected). The action space $\mathcal{A} = \{U, D\}$ consists of the actions U (going out) and D (keeping distance). The horizon is $H = 50$. The reward functions are defined as $r_h^*(s, a, z) = -10 \cdot \mathbb{I}_{s=I} - 2.5 \cdot \mathbb{I}_{s=D}$ for all $h \in [H]$. The transition kernels are defined as

$$P_h^*(S \mid I, \cdot, \cdot) = 0.2, \quad P_h^*(I \mid S, U, z) = 0.8 \cdot z(I), \quad P_h^*(I \mid S, D, \cdot) = 0 \text{ for all } h \in [H].$$

Investment problem: The state space is $\mathcal{S} = \{0, 1, \ldots, 9\}$, and the action space is $\mathcal{A} = \{I, O\}$. The horizon is $H = 50$. The reward function is

$$r_h^*(s, a, z) = \frac{0.3 \cdot s}{1 + \sum_{s' \in \mathcal{S}} s' * z(s')} - 2 \cdot \mathbb{I}_{a=I}$$

for all $h \in [H]$. The transition kernel is specified as

$$P_h^*(s+1 \mid s, I, \cdot) = \frac{9-s}{10}, \quad P_h^*(s \mid s, I, \cdot) = \frac{1+s}{10}, \quad P_h^*(s \mid s, O, \cdot) = 1$$

for all $s \in \{0, \ldots, 8\}$, and $s = 9$ is an absorbing state.

We then introduce our graphon parameters. We set $\theta = 3$ for exp-graphon. For SBM graphon, we set $K = 2$, $p_0 = 1, p_1 = 0.7, p_2 = 1$, $a_{11} = a_{22} = 0.9$, and $a_{12} = a_{21} = 0.3$. We set $a = 1, b = 0.5$ for affine attachment graphon and ranked attachment graphon. We set the regularization parameter as $\lambda = 1$ in our experiments. For the choices of the model classes, we note that the SIS and investment problems involve a set of parameters. For example, the coefficients 10 and 2.5 for the reward function of SIS problem. We estimate these coefficients. For the graphon classes, we note that all the graphons can be parameterized by some parameters, and we estimate these parameters in the experiments. For the implementation of $\pi^{\mathcal{I}}$ and the computation of $\mu^{\mathcal{I}}$, we discretize the infinitely many agents indexed by $[0, 1]$ into $N = 100$ groups, and approximate the policies and distribution flows within each group by one policy and one distribution flow. This step incurs an approximation error $O(N^{-1})$ with respect to the ℓ_1 norm.

To shorten the simulation time and convey the main message, we only estimate the model in the beginning of the first iteration round and reuse this estimate in the following iterations to generate action-value function estimates. Figures 2, 3 and 4 are derived from twenty Monte-Carlo implementations of the algorithms. The error bar indicates the 25% and the 75% quantile of the errors. When simulating the cases with constant graphons, we implement the fixed point iteration of the mirror descent operator (Yardim et al., 2022; Xie et al., 2021) or the game operator (Guo et al., 2019) to find the NE, and the calculations of the optimal policy and the induced distribution flow are implemented via the dynamical programming and direct calculation with the nominal transition kernels and reward functions. Thus, there is no error bar for these cases.

The code used in our simulations uses the code in Fabian et al. (2022) and Cui and Koeppl (2021b) for building the simulation environment. We run our simulations on Intel(R) Core(TM) i5-8257U CPU @ 1.40GHz.

Appendix B. Discussion of Regularization

We focus on regularized GMFGs in the paper. Here, for the sake of completeness, we discuss the relationship between regularized and unregularized games. We show that the NE of the regularized GMFG suffers at most a $\lambda H \log|\mathcal{A}|$ exploitability compared to that of unregularized MFGs. For a policy $\pi^{\mathcal{I}}$, we denote its *exploitability* in a λ-regularized GMFG as

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

We note that the exploitability $\Delta(\pi^{\mathcal{I}})$ defined in Section 7 is indeed $\Delta^{\lambda}(\pi^{\mathcal{I}})$ here. Then Proposition 3 in Geist et al. (2019) asserts that

$$\big|\Delta^{\lambda}(\pi^{\mathcal{I}}) - \Delta^{0}(\pi^{\mathcal{I}})\big| \le \lambda H \log|\mathcal{A}|$$

for all $\lambda \ge 0$. This inequality implies that the NE of the regularized GMFG (resp. unregularized) satisfies agent rationality in the unregularized (resp. regularized) up to $\lambda H \log|\mathcal{A}|$. This gap also appears in MFGs (Anahtarci et al., 2022; Xie et al., 2021), and mitigating the bias remains an unsolved problem in MFGs, a strict subclass of GMFGs.

Appendix C. Corollary of Theorem 5 in Single-policy Setting

We next derive a corollary for the setting where we implement a single behavior policy for L independent times to collect the data, i.e., $\pi^{\mathcal{I}}_{\tau} = \pi^{\mathcal{I}}$ for all $\tau \in [L]$. As such, instead of Eqn. (4), we estimate the mean-embedding via

$$\hat{\hat{\omega}}^{i}_{\tau,h}(W) = \frac{1}{(N-1)L}\sum_{j\neq i}\sum_{\tau'=1}^{L} W(\xi_i,\xi_j)k\big(\cdot,(s^i_{\tau,h}, a^i_{\tau,h}, s^j_{\tau',h})\big). \tag{14}$$

We note that Eqn. (14) averages the states over L episodes, since the distribution flows of these L episodes are same. Correspondingly, the estimation procedure in Eqn. (5) is modified to be

$$(\hat{f}_h, \hat{g}_h, \hat{W}_h) = \mathop{\mathrm{argmin}}_{\substack{f\in\mathbb{B}(r,\bar{\mathcal{H}}), g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}),\\ \tilde{W}\in\tilde{\mathcal{W}}}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - f\big(\hat{\hat{\omega}}^i_{\tau,h}(\tilde{W})\big)\Big)^2 + \Big(r^i_{\tau,h} - g\big(\hat{\hat{\omega}}^i_{\tau,h}(\tilde{W})\big)\Big)^2. \tag{15}$$

The convergence rate of the corresponding estimates $(\hat{f}_h, \hat{g}_h, \hat{W}_h)$ can be derived as follows.

Corollary 12 *Under Assumptions 1, 5, 6, and 7, if we implement a policy L independent times to collect the data, and $\xi_i = i/N$ for $i \in [N]$, then with probability at least $1-\delta$, the risk of the estimates in Eqn.* (15) *can be bounded as*

$$\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*) \\ = O\biggl(\frac{(B_S + \bar{r}B_K)^4}{NL} \log \frac{N_{\mathbb{B}_r} N_{\tilde{\mathbb{B}}_{\tilde{r}}} N_{\tilde{\mathcal{W}}}}{\delta} + \frac{(B_S + \bar{r}B_K)\bar{r}L_K B_k}{\sqrt{NL}} \log \frac{NL\mathcal{N}_\infty(1/\sqrt{NL}, \tilde{\mathcal{W}})}{\delta}\biggr),$$

where

$$N_{\mathbb{B}_r} = \mathcal{N}_{\bar{\mathcal{H}}}\Bigl(\frac{3}{NL}, \mathbb{B}(r, \bar{\mathcal{H}})\Bigr), \quad N_{\tilde{\mathbb{B}}_{\tilde{r}}} = \mathcal{N}_{\bar{\mathcal{H}}}\Bigl(\frac{3}{NL}, \mathbb{B}(\tilde{r}, \tilde{\mathcal{H}})\Bigr), \quad N_{\tilde{\mathcal{W}}} = \mathcal{N}_\infty\Bigl(\frac{3}{L_K NL}, \tilde{\mathcal{W}}\Bigr).$$

The proof is provided in Appendix J. Compared to the result in Theorem 5, the mean-embedding estimation error, i.e., the second term, is improved from $O(1/\sqrt{N})$ to $O(1/\sqrt{NL})$. Such an improvement is intuitive, since we now utilize the data from L episodes to estimate the distribution flow, but the estimation procedure in Theorem 5 only uses the data from a *single* episode for the same purpose.

Appendix D. Proof of Proposition 2

Proof [Proof of Proposition 2] We prove the desired results by induction on $h \in [H]$. When $h = 1$, $\mu_1^{\phi(\alpha)} = \mu_1^{\phi,\alpha}$ holds trivially for all $\alpha \in \mathcal{I}$. Assume that $\mu_h^{\phi(\alpha)} = \mu_h^{\phi,\alpha}$ holds for all $\alpha \in \mathcal{I}$, then for $h+1$ and any $\alpha \in \mathcal{I}$ we have that

$$\mu_{h+1}^{\phi(\alpha)}(s') = \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} \mu_h^{\phi(\alpha)}(s)\pi_h^{\phi(\alpha)}(a \mid s)P_h^*(s' \mid s, a, z_h^{\phi(\alpha)}(\mu_h^{\mathcal{I}}, W_h^*))\,\mathrm{d}s \qquad \text{and}$$

$$\begin{aligned}\mu_{h+1}^{\phi,\alpha}(s') &= \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} \mu_h^{\phi,\alpha}(s)\pi_h^{\phi,\alpha}(a \mid s)P_h^*(s' \mid s, a, z_h^{\alpha}(\mu_h^{\phi,\mathcal{I}}, W_h^{\phi,*}))\,\mathrm{d}s \\ &= \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} \mu_h^{\phi(\alpha)}(s)\pi_h^{\phi(\alpha)}(a \mid s)P_h^*(s' \mid s, a, z_h^{\alpha}(\mu_h^{\phi,\mathcal{I}}, W_h^{\phi,*}))\,\mathrm{d}s,\end{aligned}$$

where the last equality results from the definition of $\pi^{\phi,\mathcal{I}}$ and the hypothesis. To show that $\mu_{h+1}^{\phi(\alpha)} = \mu_{h+1}^{\phi,\alpha}$, it remains to show $z_h^{\phi(\alpha)}(\mu_h^{\mathcal{I}}, W_h^*) = z_h^{\alpha}(\mu_h^{\phi,\mathcal{I}}, W_h^{\phi,*})$. In fact, we have that

$$z_h^{\alpha}(\mu_h^{\phi,\mathcal{I}}, W_h^{\phi,*}) = \int_0^1 W_h^*(\phi(\alpha), \phi(\beta))\mu_h^{\phi(\beta)}\,\mathrm{d}\beta = \int_0^1 W_h^*(\phi(\alpha), \gamma)\mu_h^{\gamma}\,\mathrm{d}\gamma,$$

where the last equality results from setting $\gamma = \phi(\beta)$. Thus, we conclude the proof of Proposition 2. ■

Appendix E. Proof of Theorem 4

Proof [Proof of Theorem 4] For the analysis of the Algorithm 1, we define the nominal distribution flows as

$$\mu_t^{\mathcal{I}} = \Gamma_2(\pi_t^{\mathcal{I}}, W^*), \quad \bar{\mu}_{t+1}^{\mathcal{I}} = (1-\alpha_t)\bar{\mu}_t^{\mathcal{I}} + \alpha_t\mu_t^{\mathcal{I}} \text{ for all } t \in [T].$$

Our proof of Theorem 4 involves four distinct steps:

- First, we derive the first-order optimality condition of the policy $\hat{\pi}^{\mathcal{I}}_{t+1}$ derived in Line 6 of Algorithm 1.
- Second, we derive the recurrence relationship of the policy learning error from the relationship in the second step.
- Third, we derive the convergence rate of the learned mean-field.
- Finally, we obtain the desired result by combining step 2 and step 3.

Step 1: Analyze the property of the policy $\hat{\pi}^{\mathcal{I}}_{t+1}$ derived in Line 6 of Algorithm 1

We first note that the update of $\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s)$ in Line 6 of Algorithm 1 can be equivalently defined as

$$\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s) = \operatorname*{argmax}_{p\in\Delta(\mathcal{A})} \eta_{t+1}\Big[\big\langle \hat{Q}^{\lambda,\alpha}_h(s,\cdot,\pi^{\alpha}_t,\hat{\bar{\mu}}^{\mathcal{I}}_t,\hat{W}),p\big\rangle - \lambda\bar{H}(p)\Big] - \mathrm{KL}\big(p\|\pi^{\alpha}_{t,h}(\cdot\,|\,s)\big), \tag{16}$$

which can be proved using Lagrangian multipliers.

Proposition 13 *For the policy $\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s)$, which is defined in Eqn.* (16)*, we have that for all $s\in\mathcal{S}$, $p\in\Delta(\mathcal{A})$, and $h\in[H]$*

$$\begin{aligned}&\eta_{t+1}\big\langle \hat{Q}^{\lambda,\alpha}_h(s,\cdot,\pi^{\alpha}_t,\hat{\bar{\mu}}^{\mathcal{I}}_t,\hat{W}),p-\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s)\big\rangle + \lambda\eta_{t+1}\Big[R\big(\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s)\big)-\bar{H}(p)\Big]\\ &\quad\le \mathrm{KL}\big(p\|\pi^{\alpha}_{t,h}(\cdot\,|\,s)\big) - (1+\lambda\eta_{t+1})\mathrm{KL}\big(p\|\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s)\big) - \mathrm{KL}\big(\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s)\|\pi^{\alpha}_{t,h}(\cdot\,|\,s)\big).\end{aligned}$$

Proof [Proof of Proposition 13] See Appendix Q.2.1. ■

Proposition 13 shows that

$$\begin{aligned}&\eta_{t+1}\big\langle Q^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\bar{\mu}^{\mathcal{I}}_t,W^*),p-\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big\rangle + \lambda\eta_{t+1}\Big[R\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big)-\bar{H}(p)\Big]\\ &\quad\le \mathrm{KL}\big(p\|\pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big) - (1+\lambda\eta_{t+1})\mathrm{KL}\big(p\|\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big) - \mathrm{KL}\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\|\pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big)\\ &\qquad + \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)},\end{aligned} \tag{17}$$

where the term (I) is the combination of the action-value function estimation error and the difference between $\hat{\pi}^{\mathcal{I}}_{t+1}$ and $\pi^{\mathcal{I}}_{t+1}$ that is defined as

$$\text{(I)} = \eta_{t+1}\Big[\big\langle Q^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\bar{\mu}^{\mathcal{I}}_t,W^*),p-\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big\rangle - \big\langle \hat{Q}^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\hat{\bar{\mu}}^{\mathcal{I}}_t,\hat{W}),p-\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big\rangle\Big].$$

The term (II) is the entropy difference between $\hat{\pi}^{\mathcal{I}}_{t+1}$ and $\pi^{\mathcal{I}}_{t+1}$ that is defined as

$$\text{(II)} = \lambda\eta_{t+1}\Big(R\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big) - R\big(\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big)\Big).$$

The term (III) is the KL divergence difference between $\hat{\pi}^{\mathcal{I}}_{t+1}$ and $\pi^{\mathcal{I}}_{t+1}$ that is defined as

$$\text{(III)} = \mathrm{KL}\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\|\pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big) - \mathrm{KL}\big(\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\|\pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big).$$

The term (IV) is also the KL divergence difference between $\hat{\pi}^{\mathcal{I}}_{t+1}$ and $\pi^{\mathcal{I}}_{t+1}$ that is defined as

$$(\mathrm{IV}) = (1+\lambda\eta_{t+1})\Big[\mathrm{KL}\big(p\|\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big) - \mathrm{KL}\big(p\|\hat{\pi}^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big)\Big].$$

We define

$$\begin{aligned}\Lambda^{\alpha}_{t+1,h} &= 2\eta_{t+1}\big\|Q^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\hat{\bar{\mu}}^{\mathcal{I}}_t,W^*) - \hat{Q}^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\hat{\bar{\mu}}^{\mathcal{I}}_t,\hat{W})\big\|_\infty + 2\eta_{t+1}H(1+\lambda\log|\mathcal{A}|)\beta_{t+1}\\ &\quad + 2\eta_{t+1}\big[L_r + H(1+\lambda\log|\mathcal{A}|)L_P\big]\varepsilon_\mu + 2\beta_{t+1}\log\frac{|\mathcal{A}|}{\beta_t} + 2(1+\lambda\eta_{t+1})\beta_{t+1}, \quad (18)\end{aligned}$$

Then we can show the following bound.

Proposition 14 *Under assumptions in Theorem 4, (I) + (II) + (III) + (IV)* $\leq \Lambda^{\alpha}_{t+1,h}$.

Proof See Appendix Q.2.2. ■

Then inequality (17) shows that

$$\begin{aligned}&\eta_{t+1}\big\langle Q^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\bar{\mu}^{\mathcal{I}}_t,W^*), p - \pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big\rangle + \lambda\eta_{t+1}\Big[R\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big) - \bar{H}(p)\Big]\\ &\qquad + \mathrm{KL}\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\|\pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big)\\ &\quad\leq \mathrm{KL}\big(p\|\pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big) - (1+\lambda\eta_{t+1})\mathrm{KL}\big(p\|\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big) + \Lambda^{\alpha}_{t+1,h}. \quad (19)\end{aligned}$$

Step 2: Derive the recurrence relationship of the policy learning error from the relationship the second step, and bound the dynamical error in such recurrence relationship.

Inequality (19) implies that the improvement of $\pi^{\mathcal{I}}_{t+1}$ of the MDP induced by $\bar{\mu}^{\mathcal{I}}_t$ over $\pi^{\mathcal{I}}_t$ can be lower bounded as

$$\begin{aligned}&V^{\lambda,\alpha}_m(s,\pi^{\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_t,W^*) - V^{\lambda,\alpha}_m(s,\pi^{\alpha}_t,\bar{\mu}^{\mathcal{I}}_t,W^*)\\ &\quad= \mathbb{E}_{\pi^{\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_t}\Bigg[\sum_{h=m}^{H}\big\langle Q^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\bar{\mu}^{\mathcal{I}}_t,W^*), \pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h) - \pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big\rangle\\ &\qquad + \lambda\Big[R\big(\pi^{\alpha}_{t,h}(\cdot\,|\,s_h)\big) - R\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big)\Big]\,\Bigg|\,s_m = s\Bigg]\\ &\quad\geq \big\langle Q^{\lambda,\alpha}_m(s,\cdot,\pi^{\alpha}_t,\bar{\mu}^{\mathcal{I}}_t,W^*), \pi^{\alpha}_{t+1,m}(\cdot\,|\,s) - \pi^{\alpha}_{t,m}(\cdot\,|\,s)\big\rangle + \lambda\Big[R\big(\pi^{\alpha}_{t,h}(\cdot\,|\,s)\big) - R\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s)\big)\Big]\\ &\qquad - \frac{1}{\eta_{t+1}}\mathbb{E}_{\pi^{\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_t}\Bigg[\sum_{h=m}^{H}\Lambda^{\alpha}_{t+1,h}\,\Bigg|\,s_m = s\Bigg], \quad (20)\end{aligned}$$

where the equality results from Lemma 37, and the inequality results from inequality (19) and that KL divergence is non-negative.

We denote the optimal policy on the MDP induced by $\bar{\mu}^{\mathcal{I}}_t$ as $\bar{\pi}^{*,\mathcal{I}}_t = \Gamma^{\lambda}_1(\bar{\mu}^{\mathcal{I}}_t, W^*)$. Then Lemma 37 and inequality (20) implies that

$$\begin{aligned}&\eta_{t+1}\mathbb{E}_{\bar{\pi}^{*,\alpha}_t,\bar{\mu}^{\mathcal{I}}_t}\Bigg[\sum_{h=1}^{H}\big\langle Q^{\lambda,\alpha}_h(s_h,\cdot,\pi^{\alpha}_t,\bar{\mu}^{\mathcal{I}}_t,W^*), \bar{\pi}^{*,\alpha}_{t,h}(\cdot\,|\,s_h) - \pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big\rangle\\ &\qquad\qquad + \lambda\Big[R\big(\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big) - R\big(\bar{\pi}^{*,\alpha}_{t,h}(\cdot\,|\,s_h)\big)\Big]\Bigg]\end{aligned}$$

$$
\begin{aligned}
&\geq \eta_{t+1}\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h,\pi_t^{\alpha},\bar{\mu}_t^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s_h,\pi_{t+1}^{\alpha},\bar{\mu}_t^{\mathcal{I}},W^*)\Bigg] \\
&\quad - \mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H}\mathbb{E}_{\pi_{t+1}^{\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{m=h}^{H}\Lambda_{t+1,m}^{\alpha}\,\Bigg|\, s_h\Bigg]\Bigg] \\
&\quad + \eta_{t+1}\mathbb{E}_{\mu_1^{\alpha}}\big[V_1^{\lambda,\alpha}(s_1,\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}},W^*) - V_1^{\lambda,\alpha}(s_1,\pi_t^{\alpha},\bar{\mu}_t^{\mathcal{I}},W^*)\big].
\end{aligned}
\tag{21}
$$

Applying inequality (19) with $p = \bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s_h)$ to the left-hand side of inequality (21) and rearranging the terms, we have that

$$
\begin{aligned}
&\eta_{t+1}\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h,\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s_h,\pi_{t+1}^{\alpha},\bar{\mu}_t^{\mathcal{I}},W^*)\Bigg] \\
&\qquad + (1+\lambda\eta_{t+1})\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H}\mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s_h)\|\pi_{t+1,h}^{\alpha}(\cdot\,|\,s_h)\big)\Bigg] \\
&\leq \eta_{t+1}\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h,\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s_h,\pi_t^{\alpha},\bar{\mu}_t^{\mathcal{I}},W^*)\Bigg] \\
&\qquad - \eta_{t+1}\mathbb{E}_{\mu_1^{\alpha}}\big[V_1^{\lambda,\alpha}(s_1,\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}},W^*) - V_1^{\lambda,\alpha}(s_1,\pi_t^{\alpha},\bar{\mu}_t^{\mathcal{I}},W^*)\big] \\
&\qquad + \mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H}\mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s_h)\|\pi_{t,h}^{\alpha}(\cdot\,|\,s_h)\big)\Bigg] \\
&\qquad + \mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H}\mathbb{E}_{\pi_{t+1}^{\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{m=h}^{H}\Lambda_{t+1,m}^{\alpha}\,\Bigg|\, s_h\Bigg]\Bigg] + \mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H}\Lambda_{t+1,h}^{\alpha}\Bigg].
\end{aligned}
\tag{22}
$$

To handle the right-hand side of this inequality, we utilize the following proposition.

Proposition 15 *For a λ-regularized finite-horizon MDP $(\mathcal{S},\mathcal{A},H,\{r_h\}_{h=1}^{H},\{P_h\}_{h=1}^{H})$ with $|r_h|\leq 1$ for all $h\in[H]$, we denote the optimal policy as $\pi^* = \{\pi_h^*\}_{h=1}^{H}$. Then for any policy π, we have that*

$$
\mathbb{E}_{\pi^*}\big[V_1^{\lambda}(s_1,\pi^*) - V_1^{\lambda}(s_1,\pi)\big] \geq \beta^*\mathbb{E}_{\pi^*}\Bigg[\sum_{h=2}^{H} V_h^{\lambda}(s_h,\pi^*) - V_h^{\lambda}(s_h,\pi)\Bigg],
$$

where the expectation is taken with respect to the state distribution induced by π^*, and $\beta^*>0$ is a constant that only depends on λ, H and $|\mathcal{A}|$.

Proof [Proof of Proposition 15] See Appendix Q.2.3. ■

Define $\theta^* = 1/(1+\beta^*) < 1$ and let $\eta_t = \eta$, where $1+\lambda\eta = 1/\theta^*$. Proposition 15 shows that

$$
\begin{aligned}
&\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h,\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s_h,\pi_{t+1}^{\alpha},\bar{\mu}_t^{\mathcal{I}},W^*)\Bigg] \\
&\qquad + \frac{1}{\eta\theta^*}\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^{H}\mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s_h)\|\pi_{t+1,h}^{\alpha}(\cdot\,|\,s_h)\big)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \theta^* \bigg\{ \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H V_h^{\lambda,\alpha}(s_h, \bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}, W^*) - V_h^{\lambda,\alpha}(s_h, \pi_t^{\alpha}, \bar{\mu}_t^{\mathcal{I}}, W^*) \bigg] \\
&\quad + \frac{1}{\eta\theta^*} \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H \mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot \,|\, s_h) \| \pi_{t,h}^{\alpha}(\cdot \,|\, s_h)\big) \bigg] \bigg\} \\
&\quad + \frac{1}{\eta} \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H \mathbb{E}_{\pi_{t+1}^{\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{m=h}^H \Lambda_{t+1,m}^{\alpha} \,\bigg|\, s_h \bigg] \bigg] \frac{1}{\eta} \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H \Lambda_{t+1,h}^{\alpha} \bigg]. \qquad (23)
\end{aligned}
$$

In the following, we will derive the rate of convergence of the following term

$$
\begin{aligned}
X_t^{\alpha} = \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H V_h^{\lambda,\alpha}(s_h, \bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}, W^*) - V_h^{\lambda,\alpha}(s_h, \pi_t^{\alpha}, \bar{\mu}_t^{\mathcal{I}}, W^*) \bigg] \\
+ \frac{1}{\eta\theta^*} \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H \mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot \,|\, s_h) \| \pi_{t,h}^{\alpha}(\cdot \,|\, s_h)\big) \bigg]. \qquad (24)
\end{aligned}
$$

We note that $X_t^{\mathcal{I}}$ is a good quantity to measure the "distance" between $\pi_t^{\mathcal{I}}$ and NE. For NE, $\pi^{*,\mathcal{I}}$ is the optimal policy on the MDP induced by the distribution flow $\mu^{*,\mathcal{I}}$ of itself. Since $\bar{\mu}_t^{\mathcal{I}}$ is close to $\mu_t^{\mathcal{I}}$, we expect that $\pi_t^{\mathcal{I}}$ achieves high rewards on the MDP induced by $\bar{\mu}_t^{\mathcal{I}}$ if it is close to the NE. Inequality (23) shows that the recurrence relationship of X_t^{α} is

$$
X_{t+1}^{\alpha} \leq \theta^* X_t^{\alpha} + \frac{1}{\eta} \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H \mathbb{E}_{\pi_{t+1}^{\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{m=h}^H \Lambda_{t+1,m}^{\alpha} \,\bigg|\, s_h \bigg] \bigg] + \frac{1}{\eta} \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H \Lambda_{t+1,h}^{\alpha} \bigg] + \Delta_{t+1}^{\alpha}, \qquad (25)
$$

where Δ_{t+1}^{α} is the error introduced by the change of the environment, which is also called the dynamical error, and it is defined as

$$
\begin{aligned}
\Delta_{t+1}^{\alpha} = X_{t+1}^{\alpha} - \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H V_h^{\lambda,\alpha}(s_h, \bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}, W^*) - V_h^{\lambda,\alpha}(s_h, \pi_{t+1}^{\alpha}, \bar{\mu}_t^{\mathcal{I}}, W^*) \bigg] \\
- \frac{1}{\eta\theta^*} \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}} \bigg[\sum_{h=1}^H \mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot \,|\, s_h) \| \pi_{t+1,h}^{\alpha}(\cdot \,|\, s_h)\big) \bigg].
\end{aligned}
$$

Proposition 16 *Under assumptions in Theorem 4, we have*

$$
\begin{aligned}
\Delta_{t+1}^{\alpha} \leq &\bigg[H \bigg(2H(1 + \lambda \log|\mathcal{A}|) + \lambda L_R + \frac{1}{\eta\theta^*} \log \frac{|\mathcal{A}|^2}{\beta_{t+1}} \bigg) + \frac{2}{\eta\theta^*} \max\bigg\{ \log \frac{|\mathcal{A}|}{\beta_{t+1}}, L_R \bigg\} \bigg] \\
&\cdot \mathbb{E}_{\bar{\pi}_{t+1}^{*,\alpha}, \bar{\mu}_{t+1}^{\mathcal{I}}} \bigg[\sum_{m=1}^H \big\| \bar{\pi}_{t+1,m}^{*,\alpha}(\cdot \,|\, s_m) - \bar{\pi}_{t,m}^{*,\alpha}(\cdot \,|\, s_m) \big\|_1 \bigg] \\
&+ \bigg[H \bigg(H\big(1 + \lambda \log|\mathcal{A}|\big) + \frac{1}{\eta\theta^*} \log \frac{|\mathcal{A}|^2}{\beta_{t+1}} \bigg) L_P + 2H\big[L_r + H(1 + \lambda \log|\mathcal{A}|) L_P \big] \bigg]
\end{aligned}
$$

$$\cdot \sum_{m=1}^{H} \int_0^1 \|\bar{\mu}_{t+1,m}^{\beta} - \bar{\mu}_{t,m}^{\beta}\|_1 \mathrm{d}\beta$$

$$= C_1(\eta, \beta_{t+1}) \mathbb{E}_{\bar{\pi}_{t+1}^{*,\alpha}, \bar{\mu}_{t+1}^{\mathcal{I}}} \Bigg[\sum_{m=1}^{H} \big\| \bar{\pi}_{t+1,m}^{*,\alpha}(\cdot \,|\, s_m) - \bar{\pi}_{t,m}^{*,\alpha}(\cdot \,|\, s_m) \big\|_1 \Bigg]$$

$$+ C_2(\eta, \beta_{t+1}) \sum_{m=1}^{H} \int_0^1 \|\bar{\mu}_{t+1,m}^{\beta} - \bar{\mu}_{t,m}^{\beta}\|_1 \mathrm{d}\beta.$$

In the above, we defined $C_1(\eta, \beta_{t+1})$ and $C_2(\eta, \beta_{t+1})$ for ease of notation subsequently.

Proof See Appendix Q.2.4. ■

We need the following proposition to relate the difference between the optimal policies $\bar{\pi}_{t+1,m}^{*,\alpha}(\cdot \,|\, s_m)$ and $\bar{\pi}_{t,m}^{*,\alpha}(\cdot \,|\, s_m)$ in Proposition 16 to the distribution flows $\bar{\mu}_{t+1}^{\mathcal{I}}$ and $\bar{\mu}_{t}^{\mathcal{I}}$.

Proposition 17 *For any two distribution flows $\mu^{\mathcal{I}}$ and $\tilde{\mu}^{\mathcal{I}}$, we define the optimal policies $\pi^{*,\mathcal{I}} = \Gamma_1^{\lambda}(\mu^{\mathcal{I}}, W^*)$ and $\tilde{\pi}^{*,\mathcal{I}} = \Gamma_1^{\lambda}(\tilde{\mu}^{\mathcal{I}}, W^*)$. Under Assumption 1, we have that for any $h \in [H]$ and $\alpha \in [0, 1]$*

$$\max_{s \in \mathcal{S}} \big| V_h^{\lambda,\alpha}(s, \pi^{*,\mathcal{I}}, \mu^{\mathcal{I}}, W^*) - V_h^{\lambda,\alpha}(s, \tilde{\pi}^{*,\mathcal{I}}, \tilde{\mu}^{\mathcal{I}}, W^*) \big|$$

$$\leq \big(H(1 + \lambda \log |\mathcal{A}|) L_P + L_r \big) \sum_{m=h}^{H} \int_0^1 \|\mu_m^{\beta} - \tilde{\mu}_m^{\beta}\|_1 \mathrm{d}\beta,$$

$$\max_{s \in \mathcal{S}} \big\| \pi_h^{*,\alpha}(\cdot \,|\, s) - \tilde{\pi}_h^{*,\alpha}(\cdot \,|\, s) \big\|_1$$

$$\leq 2 \big(H(1 + \lambda \log |\mathcal{A}|) L_P + L_r \big) \sum_{m=h}^{H} \int_0^1 \|\mu_m^{\beta} - \tilde{\mu}_m^{\beta}\|_1 \mathrm{d}\beta.$$

Proof [Proof of Proposition 17] See Appendix Q.2.9. ■

Propositions 16 and 17 shows that

$$\Delta_{t+1}^{\alpha} \leq \Big(2H \big(H(1 + \lambda \log |\mathcal{A}|) L_P + L_r \big) C_1(\eta, \beta_{t+1}) + C_2(\eta, \beta_{t+1}) \Big) \sum_{m=1}^{H} \int_0^1 \|\bar{\mu}_{t+1,m}^{\beta} - \bar{\mu}_{t,m}^{\beta}\|_1 \mathrm{d}\beta$$

$$\leq 2H \Big(2H \big(H(1 + \lambda \log |\mathcal{A}|) L_P + L_r \big) C_1(\eta, \beta_{t+1}) + C_2(\eta, \beta_{t+1}) \Big) \alpha_t, \tag{26}$$

where the inequality results from the definition of $\bar{\mu}_{t+1}^{\mathcal{I}}$.

Next, we will combine Eqn. (18) and inequalities (25) and (26) to derive a relationship between X_{t+1}^{α} and X_t^{α}. Adopting Assumption 4 to control the estimation error of the action-value functions in inequality (25), we have that

$$\int_0^1 X_{t+1}^{\alpha} \mathrm{d}\alpha \leq \theta^* \int_0^1 X_t^{\alpha} \mathrm{d}\alpha + \frac{H(H+1)}{\eta} \Bigg[2\eta \varepsilon_Q + 2\eta H(1 + \lambda \log |\mathcal{A}|) \beta_{t+1} + 2\beta_{t+1} \log \frac{|\mathcal{A}|}{\beta_t}$$

$$+ 2\eta \big[L_r + H(1 + \lambda \log |\mathcal{A}|) L_P \big] \varepsilon_{\mu} + 2(1 + \lambda\eta) \beta_{t+1} \Bigg]$$

$$+ 2H\Big(2H\big(H(1+\lambda\log|\mathcal{A}|)L_P + L_r\big)C_1(\eta,\beta_{t+1}) + C_2(\eta,\beta_{t+1})\Big)\alpha_t, \quad (27)$$

where the inequality results from Assumption 4.

We set $\alpha_t = O(T^{-2/3})$ and $\beta_t = O(T^{-1})$ for all $t \in [T]$. Lemma 40 shows that

$$\int_0^1 X_t^\alpha \mathrm{d}\alpha = O\bigg((\theta^*)^t + (\varepsilon_Q + \varepsilon_\mu)(\theta^*)^{t/2} + \frac{\log T}{T^{2/3}}\bigg) + O(\varepsilon_Q + \varepsilon_\mu).$$

Thus, we have

$$\frac{1}{T}\sum_{t=1}^T \int_0^1 X_t^\alpha \mathrm{d}\alpha = O\bigg(\frac{\log T}{T^{2/3}}\bigg) + O(\varepsilon_Q + \varepsilon_\mu). \quad (28)$$

Step 3: Derive the convergence rate of the learned mean-field.

To derive the convergence behavior of $\hat{\bar{\mu}}_t^{\mathcal{I}}$, we define the distribution flow induced by $\bar{\pi}_t^{*,\mathcal{I}}$ as $\bar{\mu}_t^{*,\mathcal{I}} = \Gamma_2(\bar{\pi}_t^{*,\mathcal{I}}, W^*)$. Then we have that

$$\begin{aligned} d(\hat{\bar{\mu}}_{t+1}^{\mathcal{I}}, \mu^{*,\mathcal{I}}) &= d\big((1-\alpha_t)\hat{\bar{\mu}}_t^{\mathcal{I}} + \alpha_t\hat{\mu}_t^{\mathcal{I}}, \mu^{*,\mathcal{I}}\big) \\ &\le (1-\alpha_t)d(\hat{\bar{\mu}}_t^{\mathcal{I}}, \mu^{*,\mathcal{I}}) + \alpha_t d(\hat{\mu}_t^{\mathcal{I}}, \mu_t^{\mathcal{I}}) + \alpha_t d(\mu_t^{\mathcal{I}}, \bar{\mu}_t^{*,\mathcal{I}}) + \alpha_t d(\bar{\mu}_t^{*,\mathcal{I}}, \mu^{*,\mathcal{I}}), \quad (29) \end{aligned}$$

where the equality results from the definition of $\hat{\bar{\mu}}_{t+1}^{\mathcal{I}}$, and the inequality results from the triangle inequality. For the fourth term in the right-hand side of inequality (29), we have that

$$\begin{aligned} d(\bar{\mu}_t^{*,\mathcal{I}}, \mu^{*,\mathcal{I}}) &= d\Big(\Gamma_2\big(\Gamma_1^\lambda(\bar{\mu}_t^{\mathcal{I}}, W^*), W^*\big), \Gamma_2\big(\Gamma_1^\lambda(\mu^{*,\mathcal{I}}, W^*), W^*\big)\Big) \\ &\le d_1 d_2 d(\bar{\mu}_t^{\mathcal{I}}, \mu^{*,\mathcal{I}}) \\ &\le d_1 d_2\big(d(\bar{\mu}_t^{\mathcal{I}}, \hat{\bar{\mu}}_t^{\mathcal{I}}) + d(\hat{\bar{\mu}}_t^{\mathcal{I}}, \mu^{*,\mathcal{I}})\big), \quad (30) \end{aligned}$$

where the equality results from the definitions of $\bar{\mu}_t^{*,\mathcal{I}}$ and $\mu^{*,\mathcal{I}}$, the first inequality results from Assumption 2, and the last inequality results from the triangle inequality. We then define $\tilde{\mu}_t^{*,\mathcal{I}} = \Gamma_3(\bar{\pi}_t^{*,\mathcal{I}}, \bar{\mu}_t^{\mathcal{I}}, W^*)$. For the third term in the right-hand side of inequality (29), we have that

$$\begin{aligned} d(\mu_t^{\mathcal{I}}, \bar{\mu}_t^{*,\mathcal{I}}) &= d\big(\Gamma_2(\pi_t^{\mathcal{I}}, W^*), \Gamma_2(\bar{\pi}_t^{*,\mathcal{I}}, W^*)\big) \\ &\le d_2\int_0^1 \sum_{h=1}^H \mathbb{E}_{\mu_h^{*,\alpha}}\Big[\big\|\pi_{t,h}^\alpha(\cdot\,|\,s) - \bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s)\big\|_1\Big]\mathrm{d}\alpha \\ &= d_2\int_0^1 \sum_{h=1}^H \mathbb{E}_{\tilde{\mu}_{t,h}^{*,\alpha}}\bigg[\frac{\mu_h^{*,\alpha}(s)}{\tilde{\mu}_{t,h}^{*,\alpha}(s)}\big\|\pi_{t,h}^\alpha(\cdot\,|\,s) - \bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s)\big\|_1\bigg]\mathrm{d}\alpha \\ &\le d_2 C_\mu\sqrt{H}\sqrt{2\int_0^1 \sum_{h=1}^H \mathbb{E}_{\tilde{\mu}_{t,h}^{*,\alpha}}\Big[\mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s)\|\pi_{t,h}^\alpha(\cdot\,|\,s)\big)\Big]\mathrm{d}\alpha}, \quad (31) \end{aligned}$$

where the first inequality results from Assumption 2, and the second inequality results from Assumption 3 and the Cauchy–Schwarz inequality. Define $Y_t = d(\hat{\bar{\mu}}_t^{\mathcal{I}}, \mu^{*,\mathcal{I}})$. Combining inequalities (29), (30), and (31), we have that

$$Y_{t+1} \leq (1-\alpha_t \bar{d})Y_t + \alpha_t d(\hat{\mu}_t^{\mathcal{I}}, \mu_t^{\mathcal{I}}) + \alpha_t d_1 d_2 d(\bar{\mu}_t^{\mathcal{I}}, \hat{\mu}_t^{\mathcal{I}}) + \alpha_t d_2 C_\mu \sqrt{H}\sqrt{2\eta\theta^* \int_0^1 X_t^\alpha \mathrm{d}\alpha},$$

where $\bar{d} = 1 - d_1 d_2$.

Recall the expressions of $\bar{\mu}_t^{\mathcal{I}}$ and $\hat{\mu}_t^{\mathcal{I}}$ in Eqn. (79), we have that

$$d(\bar{\mu}_t^{\mathcal{I}}, \hat{\mu}_t^{\mathcal{I}}) \leq \sum_{m=1}^{t-1} \alpha_{m,t-1} d(\hat{\mu}_m^{\mathcal{I}}, \mu_m^{\mathcal{I}}) \leq \varepsilon_\mu,$$

where the first inequality results from the triangle inequality, and the second inequality results from Assumption 4. Take $\alpha_t = \alpha$. we have that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T Y_t &\leq \frac{1}{\bar{d}\alpha T} Y_1 + \frac{(1+d_1 d_2)}{\bar{d}}\varepsilon_\mu + \frac{d_2 C_\mu \sqrt{H}}{\bar{d}} \cdot \frac{1}{T}\sum_{t=1}^T \sqrt{2\eta\theta^* \int_0^1 X_t^\alpha \mathrm{d}\alpha} \\
&\leq \frac{1}{\bar{d}\alpha T} Y_1 + \frac{(1+d_1 d_2)}{\bar{d}}\varepsilon_\mu + \frac{d_2 C_\mu \sqrt{H}}{\bar{d}} \cdot \sqrt{2\eta\theta^* \frac{1}{T}\sum_{t=1}^T \int_0^1 X_t^\alpha \mathrm{d}\alpha},
\end{aligned}$$

where the last inequality results from Eqn. (28). Thus, we have

$$\frac{1}{T}\sum_{t=1}^T Y_t = O\biggl(\frac{\sqrt{\log T}}{T^{1/3}}\biggr) + O(\varepsilon_\mu + \sqrt{\varepsilon_Q + \varepsilon_\mu}).$$

Step 4: Build the desired result from step 2 and step 3.

From the definition of X_t, i.e., Eqn. (24), and Eqn. (28), we have that

$$\frac{1}{T}\sum_{t=1}^T \int_0^1 \mathbb{E}_{\bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}}\biggl[\sum_{h=1}^H \mathrm{KL}\bigl(\bar{\pi}_{t,h}^{*,\alpha}(\cdot \mid s_h) \| \pi_{t,h}^\alpha(\cdot \mid s_h)\bigr)\biggr]\mathrm{d}\alpha = O\biggl(\frac{\log T}{T^{2/3}}\biggr) + O(\varepsilon_Q + \varepsilon_\mu).$$

Recall that we defined $\tilde{\mu}_t^{*,\mathcal{I}} = \Gamma_3(\bar{\pi}_t^{*,\mathcal{I}}, \bar{\mu}_t^{\mathcal{I}}, W^*)$. Then we bound $D(\cdot,\cdot)$ as follows

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T D(\pi_t^{\mathcal{I}}, \bar{\pi}_t^{*,\mathcal{I}}) &= \frac{1}{T}\sum_{t=1}^T \int_0^1 \sum_{h=1}^H \mathbb{E}_{\tilde{\mu}_{t,h}^{*,\alpha}}\biggl[\frac{\mu_h^{*,\alpha}(s)}{\tilde{\mu}_{t,h}^{*,\alpha}(s)} \bigl\|\pi_{t,h}^\alpha(\cdot \mid s) - \bar{\pi}_{t,h}^{*,\alpha}(\cdot \mid s)\bigr\|_1\biggr]\mathrm{d}\alpha \\
&\leq C_\mu \sqrt{H}\sqrt{\frac{2}{T}\sum_{t=1}^T \int_0^1 \sum_{h=1}^H \mathbb{E}_{\tilde{\mu}_{t,h}^{*,\alpha}}\Bigl[\mathrm{KL}\bigl(\bar{\pi}_{t,h}^{*,\alpha}(\cdot \mid s) \| \pi_{t,h}^\alpha(\cdot \mid s)\bigr)\Bigr]\mathrm{d}\alpha} \\
&\leq O\biggl(\frac{\sqrt{\log T}}{T^{1/3}}\biggr) + O(\sqrt{\varepsilon_Q + \varepsilon_\mu}),
\end{aligned}$$

where the first inequality results from the same arguments in inequality (31). To bound the distance between $\pi_t^{\mathcal{I}}$ and $\pi^{*,\mathcal{I}}$, we adopt the triangle inequality as

$$D(\pi_t^{\mathcal{I}}, \pi^{*,\mathcal{I}}) \leq D(\pi_t^{\mathcal{I}}, \bar{\pi}_t^{*,\mathcal{I}}) + D(\bar{\pi}_t^{*,\mathcal{I}}, \pi^{*,\mathcal{I}}) \leq D(\pi_t^{\mathcal{I}}, \bar{\pi}_t^{*,\mathcal{I}}) + d_1 d(\bar{\mu}_t^{\mathcal{I}}, \mu^{*,\mathcal{I}}).$$

Thus, we have that

$$
\begin{aligned}
&D\bigg(\frac{1}{T}\sum_{t=1}^{T}\pi_t^{\mathcal{I}},\pi^{*,\mathcal{I}}\bigg)+d\bigg(\frac{1}{T}\sum_{t=1}^{T}\hat{\bar{\mu}}_t^{\mathcal{I}},\mu^{*,\mathcal{I}}\bigg)\\
&\leq \frac{1}{T}\sum_{t=1}^{T}D(\pi_t^{\mathcal{I}},\bar{\pi}_t^{*,\mathcal{I}})+d_1 d(\bar{\mu}_t^{\mathcal{I}},\mu^{*,\mathcal{I}})+d(\hat{\bar{\mu}}_t^{\mathcal{I}},\mu^{*,\mathcal{I}})\\
&\leq \frac{1}{T}\sum_{t=1}^{T}D(\pi_t^{\mathcal{I}},\bar{\pi}_t^{*,\mathcal{I}})+d_1 d(\hat{\bar{\mu}}_t^{\mathcal{I}},\mu^{*,\mathcal{I}})+d(\hat{\bar{\mu}}_t^{\mathcal{I}},\mu^{*,\mathcal{I}})+d_1 d(\bar{\mu}_t^{\mathcal{I}},\hat{\bar{\mu}}_t^{\mathcal{I}})\\
&=O\bigg(\frac{\sqrt{\log T}}{T^{1/3}}\bigg)+O(\varepsilon_\mu+\sqrt{\varepsilon_Q+\varepsilon_\mu}),
\end{aligned}
$$

where the first inequality results from Jensen's inequality, and the second inequality results from the triangle inequality. Thus, we conclude the proof of Theorem 4. ■

Appendix F. Corollary for Single-Agent MDP

In this section, we state and prove our corollary for the policy mirror descent algorithm on single-agent MDP. A single-agent MDP is defined through a tuple $(\mathcal{S},\mathcal{A},\mu_1,P,r,H)$. The state space and the action space are denoted respectively as $\mathcal{S}$ and $\mathcal{A}$. The initial state distribution $\mu_1\in\Delta(\mathcal{S})$ is state distribution at time $h=1$. The transition kernels $P_h:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S}))$ and reward functions $r_h:\mathcal{S}\times\mathcal{A}\to[0,1]$ for $h\in[H]$ describe the state transition behavior and the reward generation process. A policy $\pi=\{\pi_h\}_{h=1}^H$ is the set of mappings $\pi_h:\mathcal{S}\to\mathcal{A}$ for $h\in[H]$. Similar as the value function defined in Section 3, the value function and the action-value function of a policy π on a λ-regularized MDP are defined as

$$
\begin{aligned}
V_h^\lambda(s,\pi)&=\mathbb{E}^\pi\bigg[\sum_{t=h}^{H}r_t(s_t,a_t)-\lambda\log\pi_t(a_t\,|\,s_t)\,\bigg|\,s_h=s\bigg],\\
Q_h^\lambda(s,a,\pi)&=r_h(s,a)+\mathbb{E}\big[V_{h+1}^\lambda(s_{h+1},\pi)\,|\,s_h=s,a_h=a\big].
\end{aligned}
$$

The cumulative reward function is $J^\lambda(\pi)=\mathbb{E}_{\mu_1}[V_1^\lambda(s,\pi)]$, where the expectation is taken with respect to $s\sim\mu_1$. We denote the optimal policy as $\pi^*=\operatorname{argmax}_{\pi\in\Pi^H}J^\lambda(\pi)$. The policy mirror descent algorithm is implementing

$$
\pi_{t+1,h}(\cdot\,|\,s)\propto\big(\pi_{t,h}(\cdot\,|\,s)\big)^{1-\frac{\lambda\eta_{t+1}}{1+\lambda\eta_{t+1}}}\exp\Big(\frac{\eta_{t+1}}{1+\lambda\eta_{t+1}}Q_h^\lambda(s,\cdot,\pi_t)\Big)\text{ for all }h\in[H].
$$

for $t\in[T]$, and we set $\pi_{1,h}(\cdot\,|\,s)=\mathrm{Unif}(\mathcal{A})$ for all $s\in\mathcal{S}$. Then the convergence result of this algorithm is

Corollary 18 *Suppose that $\eta_t=\eta>0$ for all $t\in[T]$, and we set this as some function of λ, H and $|\mathcal{A}|$. Then we have*

$$
\mathbb{E}_{\pi^*}\bigg[\sum_{h=1}^{H}V_h^\lambda(s_h,\pi^*)-V_h^\lambda(s_h,\pi_{t+1})\bigg]+\frac{1}{\eta\theta^*}\mathbb{E}_{\pi^*}\bigg[\sum_{h=1}^{H}\mathrm{KL}\big(\pi_h^*(\cdot\,|\,s_h)\|\pi_{t+1,h}(\cdot\,|\,s_h)\big)\bigg]
$$

$$\leq \theta^* \bigg\{ \mathbb{E}_{\pi^*} \bigg[\sum_{h=1}^{H} V_h^\lambda(s_h, \pi^*) - V_h^\lambda(s_h, \pi_t) \bigg] + \frac{1}{\eta\theta^*} \mathbb{E}_{\pi_t^*} \bigg[\sum_{h=1}^{H} \mathrm{KL}\big(\pi_{t,h}^*(\cdot \,|\, s_h) \| \pi_{t,h}(\cdot \,|\, s_h)\big) \bigg] \bigg\},$$

where $0 < \theta^* < 1$ is a function of λ, H and $|\mathcal{A}|$, and $\mathbb{E}_{\pi^*}$ refers to the expectation with respect to the state distribution induced by π^*.

Proof [Proof of Corollary 18] Similarly as Step 1 of the proof of Theorem 4, we have

$$\eta_{t+1} \langle Q_h^\lambda(s_h, \cdot, \pi_t), p - \pi_{t+1,h}(\cdot|s_h) \rangle + \lambda\eta_{t+1} \Big[R\big(\pi_{t+1,h}(\cdot|s_h)\big) - R(p) \Big] + \mathrm{KL}\big(\pi_{t+1,h}(\cdot|s_h) \| \pi_{t,h}(\cdot|s_h)\big)$$
$$\leq \mathrm{KL}\big(p \| \pi_{t,h}(\cdot|s_h)\big) - (1 + \lambda\eta_{t+1}) \mathrm{KL}\big(p \,\|\, \pi_{t+1,h}(\cdot \,|\, s_h)\big)$$

for any $p \in \Delta(\mathcal{A})$. Following the same pipeline to inequality (23), we have that

$$\mathbb{E}_{\pi^*} \bigg[\sum_{h=1}^{H} V_h^\lambda(s_h, \pi^*) - V_h^\lambda(s_h, \pi_{t+1}) \bigg] + \frac{1}{\eta\theta^*} \mathbb{E}_{\pi^*} \bigg[\sum_{h=1}^{H} \mathrm{KL}\big(\pi_h^*(\cdot \,|\, s_h) \| \pi_{t+1,h}(\cdot \,|\, s_h)\big) \bigg]$$
$$\leq \theta^* \bigg\{ \mathbb{E}_{\pi^*} \bigg[\sum_{h=1}^{H} V_h^\lambda(s_h, \pi^*) - V_h^\lambda(s_h, \pi_t) \bigg] + \frac{1}{\eta\theta^*} \mathbb{E}_{\pi_t^*} \bigg[\sum_{h=1}^{H} \mathrm{KL}\big(\pi_{t,h}^*(\cdot \,|\, s_h) \| \pi_{t,h}(\cdot \,|\, s_h)\big) \bigg] \bigg\},$$

where β^* is defined in Proposition 15, $\theta^* = 1/(1+\beta^*) < 1$, $\eta_t = \eta$ is defined through $1 + \lambda\eta = 1/\theta^*$. We note that in this single-agent setting, we do not have the $\bar{\mu}_t^{\mathcal{I}}$, which is adopted to calculate the influence from others. Thus, the optimal policy π^* is not changed over iterations. At the same time, we do not include the estimation error in the above algorithm. We conclude the proof of Corollary 18. ■

Appendix G. Proof of Theorem 5

Proof [Proof of Theorem 5] Before delving in the formal proof, we would like to recap the definitions of covering numbers. The covering number of the graphon function class $\mathcal{N}_\infty(\varepsilon, \tilde{\mathcal{W}})$ is the minimal size C of a set of graphons $\{W_i\}_{i=1}^C$ that ε-covers $\tilde{\mathcal{W}}$. A set $\{W_i\}_{i=1}^C$ ε-covers $\tilde{\mathcal{W}}$ if for any $\tilde{W} \in \tilde{\mathcal{W}}$, there exists an index $i \in [C]$ such that $\|W_i - \tilde{W}\|_\infty \leq \varepsilon$. The covering number of the set $\mathcal{F}$ in RKHS $\bar{\mathcal{H}}$ (resp. $\tilde{\mathcal{H}}$) is defined as the minimal size C of the ε-cover set $\{f_i\}_{i=1}^C \subseteq \bar{\mathcal{H}}$. A set $\{f_i\}_{i=1}^C$ ε-covers $\mathcal{F}$ if for any $f \in \mathcal{F}$, there exists $i \in [C]$ such that $\|f - f_i\|_{\bar{\mathcal{H}}} \leq \varepsilon$ (resp. $\|f - f_i\|_{\tilde{\mathcal{H}}} \leq \varepsilon$).

For the proof of the theorem, We first decompose the difference between the risk as the sum of the generalization error of risk, the Estimation Error of mean-embedding, and the empirical risk difference. Given the fact that the empirical risk difference is equal and less to zero, Our proof involves two steps:

- Bound the Estimation Error of Mean-embedding.
- Bound the generalization error of risk.

$$\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)$$

$$
\begin{aligned}
&= \text{Generalization Error of Risk} + \text{Estimation Error of Mean-embedding} \\
&\qquad + \text{Empirical Risk Difference},
\end{aligned}
$$

where each term is defined as

$$
\begin{aligned}
&\text{Generalization Error of Risk} \\
&= \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^{i}}\left[\left(s_{\tau,h+1}^{i}-\hat{f}_h\big(\omega_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(s_{\tau,h+1}^{i}-f_h^{*}\big(\omega_{\tau,h}^{i}(W_h^{*})\big)\right)^2\right] \\
&\quad - 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(s_{\tau,h+1}^{i}-\hat{f}_h\big(\omega_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(s_{\tau,h+1}^{i}-f_h^{*}\big(\omega_{\tau,h}^{i}(W_h^{*})\big)\right)^2 \\
&\quad + \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^{i}}\left[\left(r_{\tau,h}^{i}-\hat{g}_h\big(\omega_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(r_{\tau,h}^{i}-g_h^{*}\big(\omega_{\tau,h}^{i}(W_h^{*})\big)\right)^2\right] \\
&\quad - 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(r_{\tau,h}^{i}-\hat{g}_h\big(\omega_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(r_{\tau,h}^{i}-g_h^{*}\big(\omega_{\tau,h}^{i}(W_h^{*})\big)\right)^2.
\end{aligned}
$$

This generalization error of risk represents the error due to the fact that we optimize over the empirical estimation of the risk not the population risk.

$$
\begin{aligned}
&\text{Estimation Error of Mean-embedding} \\
&= 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(s_{\tau,h+1}^{i}-\hat{f}_h\big(\omega_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(s_{\tau,h+1}^{i}-\hat{f}_h\big(\hat{\omega}_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2 \\
&\quad + 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(s_{\tau,h+1}^{i}-f_h^{*}\big(\hat{\omega}_{\tau,h}^{i}(W_h^{*})\big)\right)^2-\left(s_{\tau,h+1}^{i}-f_h^{*}\big(\omega_{\tau,h}^{i}(W_h^{*})\big)\right)^2 \\
&\quad + 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(r_{\tau,h}^{i}-\hat{g}_h\big(\omega_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(r_{\tau,h}^{i}-\hat{g}_h\big(\hat{\omega}_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2 \\
&\quad + 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(r_{\tau,h}^{i}-g_h^{*}\big(\hat{\omega}_{\tau,h}^{i}(W_h^{*})\big)\right)^2-\left(r_{\tau,h}^{i}-g_h^{*}\big(\omega_{\tau,h}^{i}(W_h^{*})\big)\right)^2.
\end{aligned}
$$

Estimation error of mean-embedding represents the error due to the fact that we cannot observe the value of $\hat{\omega}_{\tau,h}^{i}(\hat{W}_h)$. Instead, we can only estimate the value of it through the states of sampled agents.

$$
\begin{aligned}
&\text{Empirical Risk Difference} \\
&= 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(s_{\tau,h+1}^{i}-\hat{f}_h\big(\hat{\omega}_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(s_{\tau,h+1}^{i}-f_h^{*}\big(\hat{\omega}_{\tau,h}^{i}(W_h^{*})\big)\right)^2 \\
&\quad + 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\left(r_{\tau,h}^{i}-\hat{g}_h\big(\hat{\omega}_{\tau,h}^{i}(\hat{W}_h)\big)\right)^2-\left(r_{\tau,h}^{i}-g_h^{*}\big(\hat{\omega}_{\tau,h}^{i}(W_h^{*})\big)\right)^2.
\end{aligned}
$$

Empirical risk difference represents the error from that fact that we choose $(\hat{f}_h, \hat{g}_h, \hat{W}_h)$ not (f_h^*, g_h^*, W_h^*) by minimizing the empirical risk. From the procedure of Algorithm (5), we have

$$\text{Empirical Risk Difference} \leq 0.$$

Thus, we have that

$$\begin{aligned}
&\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*) \\
&\quad \leq \bigg\{ \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \mathbb{E}_{\rho^i_{\tau,h}} \Big[\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h)\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f_h^*\big(\omega^i_{\tau,h}(W_h^*)\big)\Big)^2 \Big] \\
&\qquad - 2 \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h)\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f_h^*\big(\omega^i_{\tau,h}(W_h^*)\big)\Big)^2 \\
&\qquad + \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \mathbb{E}_{\rho^i_{\tau,h}} \Big[\Big(r^i_{\tau,h} - \hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}_h)\big)\Big)^2 - \Big(r^i_{\tau,h} - g_h^*\big(\omega^i_{\tau,h}(W_h^*)\big)\Big)^2 \Big] \\
&\qquad - 2 \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(r^i_{\tau,h} - \hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}_h)\big)\Big)^2 - \Big(r^i_{\tau,h} - g_h^*\big(\omega^i_{\tau,h}(W_h^*)\big)\Big)^2 \bigg\} \\
&\qquad + 2 \sup_{f \in \mathbb{B}(r, \bar{\mathcal{H}}), \tilde{W} \in \tilde{\mathcal{W}}} \bigg| \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s^i_{\tau,h+1} - f\big(\hat{\omega}^i_{\tau,h}(\tilde{W})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f\big(\omega^i_{\tau,h}(\tilde{W})\big)\Big)^2 \bigg| \\
&\qquad + 2 \sup_{g \in \mathbb{B}(\tilde{r}, \tilde{\mathcal{H}}), \tilde{W} \in \tilde{\mathcal{W}}} \bigg| \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(r^i_{\tau,h} - g\big(\hat{\omega}^i_{\tau,h}(\tilde{W})\big)\Big)^2 - \Big(r^i_{\tau,h} - g\big(\omega^i_{\tau,h}(\tilde{W})\big)\Big)^2 \bigg| \\
&\quad = \text{(I)} + \text{(II)},
\end{aligned} \tag{32}$$

We note that the terms related to the transition kernels and reward functions are similar. In the following, we will only present the bounds for the terms related to the transition kernels, and the bounds for the reward functions can be similarly derived.

Step 1: Bound the Estimation Error of Mean-embedding.

Considering term (II), we have that

$$\begin{aligned}
&\sup_{f \in \mathbb{B}(r, \bar{\mathcal{H}}), \tilde{W} \in \tilde{\mathcal{W}}} \bigg| \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s^i_{\tau,h+1} - f\big(\hat{\omega}^i_{\tau,h}(\tilde{W})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f\big(\omega^i_{\tau,h}(\tilde{W})\big)\Big)^2 \bigg| \\
&\quad \leq \sup_{f \in \mathbb{B}(r, \bar{\mathcal{H}}), \tilde{W} \in \tilde{\mathcal{W}}} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big| f\big(\hat{\omega}^i_{\tau,h}(\tilde{W})\big) - f\big(\omega^i_{\tau,h}(\tilde{W})\big) \Big| \cdot \Big| 2 s^i_{\tau,h+1} - f\big(\hat{\omega}^i_{\tau,h}(\tilde{W}) - f\big(\omega^i_{\tau,h}(\tilde{W})\big) \Big| \\
&\quad \leq 2(B_S + r B_{\bar{K}}) r L_K \sup_{\tilde{W} \in \tilde{\mathcal{W}}} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \big\| \hat{\omega}^i_{\tau,h}(\tilde{W}) - \omega^i_{\tau,h}(\tilde{W}) \big\|_{\mathcal{H}},
\end{aligned} \tag{33}$$

where the first inequality results from the triangle inequality, and the second inequality results from Assumption 6 and Lemma 33. Recall the definitions of $\hat{\omega}^i_{\tau,h}(W)$ and $\omega^i_{\tau,h}(W)$ are

$$\omega^i_{\tau,h}(W) = \int_0^1 \int_{\mathcal{S}} W(\xi_i, \beta) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big) \mu^{\beta}_{\tau,h}(s) \, \mathrm{d}s \, \mathrm{d}\beta,$$

$$\hat{\omega}^i_{\tau,h}(W) = \frac{1}{N-1}\sum_{j\neq i} W(\xi_i,\xi_j) k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s^j_{\tau,h})\big),$$

respectively. We decompose the error between them as

$$\begin{aligned}\sup_{W\in\tilde{\mathcal{W}}}\big\|\hat{\omega}^i_{\tau,h}(W)-\omega^i_{\tau,h}(W)\big\|_{\mathcal{H}} &\leq \sup_{W\in\tilde{\mathcal{W}}}\big\|\bar{\omega}^i_{\tau,h}(W)-\omega^i_{\tau,h}(W)\big\|_{\mathcal{H}} + \sup_{W\in\tilde{\mathcal{W}}}\big\|\hat{\omega}^i_{\tau,h}(W)-\bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}}\\ &= \text{(III)}+\text{(IV)},\end{aligned} \tag{34}$$

where

$$\bar{\omega}^i_{\tau,h}(W) = \frac{1}{N-1}\sum_{j\neq i} W(\xi_i,\xi_j)\int_{\mathcal{S}} k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^j_{\tau,h}(s)\mathrm{d}s.$$

For term (III) $= \sup_{W\in\tilde{\mathcal{W}}}\big\|\omega^i_{\tau,h}(W)-\bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}}$, we have that

$$\begin{aligned}\big\|\omega^i_{\tau,h}(W)-\bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}} \leq \bigg\| &\int_0^1\int_{\mathcal{S}} W(\xi_i,\beta)k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^\beta_{\tau,h}(s)\,\mathrm{d}s\,\mathrm{d}\beta\\ &-\frac{1}{N-1}\sum_{j=1}^{N-1} W\bigg(\xi_i,\frac{j}{N-1}\bigg)\int_{\mathcal{S}} k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^{\frac{j}{N-1}}_{\tau,h}(s)\,\mathrm{d}s\bigg\|_{\mathcal{H}}\\ &+\bigg\|\frac{1}{N-1}\sum_{j=1}^{N-1} W\bigg(\xi_i,\frac{j}{N-1}\bigg)\int_{\mathcal{S}} k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^{\frac{j}{N-1}}_{\tau,h}(s)\,\mathrm{d}s\\ &-\frac{1}{N-1}\sum_{j\neq i} W(\xi_i,\xi_j)\int_{\mathcal{S}} k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^j_{\tau,h}(s)\,\mathrm{d}s\bigg\|_{\mathcal{H}}\\ = \text{(V)}+\text{(VI)}&\end{aligned} \tag{35}$$

For term (V), we have that

$$\begin{aligned}&\bigg\|\int_0^1\int_{\mathcal{S}} W(\xi_i,\beta)k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^\beta_{\tau,h}(s)\mathrm{d}s\mathrm{d}\beta\\ &\qquad-\frac{1}{N-1}\sum_{j=1}^{N-1} W\bigg(\xi_i,\frac{j}{N-1}\bigg)\int_{\mathcal{S}} k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^{\frac{j}{N-1}}_{\tau,h}(s)\,\mathrm{d}s\bigg\|_{\mathcal{H}}\\ &\leq\sum_{j=1}^{N-1}\int_{\frac{j-1}{N-1}}^{\frac{j}{N-1}}\bigg\|\int_{\mathcal{S}} W(\xi_i,\beta)k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^\beta_{\tau,h}(s)\,\mathrm{d}s\\ &\qquad-W\bigg(\xi_i,\frac{j}{N-1}\bigg)\int_{\mathcal{S}} k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^{\frac{j}{N-1}}_{\tau,h}(s)\mathrm{d}s\bigg\|_{\mathcal{H}}\mathrm{d}\beta,\end{aligned} \tag{36}$$

where the inequality results from the triangle inequality. For each term in the sum, we have that

$$\bigg\|\int_{\mathcal{S}} W(\xi_i,\beta)k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^\beta_{\tau,h}(s)\mathrm{d}s$$

$$
\begin{aligned}
&\quad - W\Big(\xi_i, \frac{j}{N-1}\Big) \int_{\mathcal{S}} k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big)\mu^{\frac{j}{N-1}}_{\tau,h}(s)\mathrm{d}s\Big\|_{\mathcal{H}} \\
&\leq \Big\|\Big(W(\xi_i, \beta) - W\Big(\xi_i, \frac{j}{N-1}\Big)\Big) \int_{\mathcal{S}} k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big)\mu^{\beta}_{\tau,h}(s)\mathrm{d}s\Big\|_{\mathcal{H}} \\
&\quad + \Big\|W\Big(\xi_i, \frac{j}{N-1}\Big) \int_{\mathcal{S}} k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big)\big(\mu^{\beta}_{\tau,h}(s) - \mu^{\frac{j}{N-1}}_{\tau,h}(s)\big)\mathrm{d}s\Big\|_{\mathcal{H}} \\
&\leq B_k L_{\mathcal{W}}\Big|\beta - \frac{j}{N-1}\Big| + B_k\Big\|\mu^{\beta}_{\tau,h} - \mu^{\frac{j}{N-1}}_{\tau,h}\Big\|_1, \qquad (37)
\end{aligned}
$$

where the first inequality results from the triangle inequality, and the second results from Assumptions 5 and 6.

Proposition 19 *Under Assumptions 5 and 1, we have that*

$$
\|\mu^{\alpha}_h - \mu^{\beta}_h\|_1 \leq (h-1)L_P L_{\mathcal{W}}|\alpha - \beta| + \sum_{t=1}^{h-1} \sup_{s\in\mathcal{S}} \big\|\pi^{\alpha}_t(\cdot\,|\,s) - \pi^{\beta}_t(\cdot\,|\,s)\big\|_1 \text{ for all } h \in [H].
$$

Proof [Proof of Proposition 19] See Appendix Q.1.1. ■

Thus, we bound the second term of inequality (37) as

$$
\begin{aligned}
\Big\|\mu^{\beta}_{\tau,h} - \mu^{\frac{j}{N-1}}_{\tau,h}\Big\|_1 &\leq HL_P L_{\mathcal{W}}\Big|\beta - \frac{j}{N-1}\Big| + \sum_{t=1}^{h-1} \sup_{s\in\mathcal{S}} \big\|\pi^{\beta}_t(\cdot\,|\,s) - \pi^{\frac{j}{N-1}}_t(\cdot\,|\,s)\big\|_1 \\
&\leq (HL_P L_{\mathcal{W}} + HL_{\pi})\Big|\beta - \frac{j}{N-1}\Big|, \qquad (38)
\end{aligned}
$$

where the first inequality results from Proposition 19, and the second inequality results from the Lipschitzness of behavior policies. Substituting inequalities (37) and (38) into inequality (36), we have that

$$
\begin{aligned}
(\mathrm{V}) &= \Big\|\int_0^1 \int_{\mathcal{S}} W(\xi_i, \beta) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big)\mu^{\beta}_{\tau,h}(s)\mathrm{d}s\mathrm{d}\beta \\
&\qquad - \frac{1}{N-1}\sum_{j=1}^{N-1} W\Big(\xi_i, \frac{j}{N-1}\Big) \int_{\mathcal{S}} k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big)\mu^{\frac{j}{N-1}}_{\tau,h}(s)\mathrm{d}s\Big\|_{\mathcal{H}} \\
&\leq \sum_{j=1}^{N-1} \int_{\frac{j-1}{N-1}}^{\frac{j}{N-1}} B_k(L_{\mathcal{W}} + HL_P L_{\mathcal{W}} + HL_{\pi})\Big|\beta - \frac{j}{N-1}\Big|\mathrm{d}\beta \\
&= \frac{1}{2(N-1)} B_k(L_{\mathcal{W}} + HL_P L_{\mathcal{W}} + HL_{\pi}).
\end{aligned}
$$

For term (VI), we have that

$$
(\mathrm{VI}) = \Big\|\frac{1}{N-1}\sum_{j=1}^{N-1} W\Big(\xi_i, \frac{j}{N-1}\Big) \int_{\mathcal{S}} k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big)\mu^{\frac{j}{N-1}}_{\tau,h}(s)\mathrm{d}s
$$

$$
\begin{aligned}
&\quad - \frac{1}{N-1}\sum_{j\neq i} W(\xi_i,\xi_j)\int_{\mathcal{S}} k\big(\cdot,(s_{\tau,h}^i,a_{\tau,h}^i,s)\big)\mu_{\tau,h}^j(s)\mathrm{d}s\Big\|_{\mathcal{H}} \\
&\leq \frac{1}{N-1}\sum_{j\neq 1}\bigg[B_k L_{\mathcal{W}}\bigg(\Big|\xi_j-\frac{j}{N-1}\Big|+\Big|\xi_j-\frac{j-1}{N-1}\Big|\bigg)+B_k\big\|\mu_{\tau,h}^{\frac{j}{N-1}}-\mu_{\tau,h}^j\big\|_1\bigg] \\
&\leq \frac{1}{N-1}\bigg[3+2B_k(L_{\mathcal{W}}+HL_\pi+HL_pL_{\mathcal{W}})\sum_{i=1}^N\Big|\xi_i-\frac{i}{N}\Big|\bigg],
\end{aligned}
$$

where the first results from triangle inequality, and the second inequality results from Proposition 19. Substituting the bounds for terms (V) and (VI) into inequality (35), we have that

$$
\begin{aligned}
\text{(III)} &= \sup_{W\in\tilde{\mathcal{W}}}\big\|\omega_{\tau,h}^i(W)-\bar{\omega}_{\tau,h}^i(W)\big\|_{\mathcal{H}} \\
&\leq \frac{B_k(L_{\mathcal{W}}+HL_PL_{\mathcal{W}}+HL_\pi)}{2(N-1)}+\frac{1}{N-1}\bigg[3+2B_k(L_{\mathcal{W}}+HL_\pi+HL_pL_{\mathcal{W}})\sum_{i=1}^N\Big|\xi_i-\frac{i}{N}\Big|\bigg] \\
&= \frac{1}{2(N-1)}B_k(L_{\mathcal{W}}+HL_PL_{\mathcal{W}}+HL_\pi)+\frac{3}{N-1}. \qquad (39)
\end{aligned}
$$

For term (IV), we have that

$$
\begin{aligned}
&\big\|\hat{\omega}_{\tau,h}^i(W)-\bar{\omega}_{\tau,h}^i(W)\big\|_{\mathcal{H}} \\
&\quad =\bigg\|\frac{1}{N-1}\sum_{j\neq i}W(\xi_i,\xi_j)k\big(\cdot,(s_{\tau,h}^i,a_{\tau,h}^i,s_{\tau,h}^j)\big)-\frac{1}{N-1}\sum_{j\neq i}W(\xi_i,\xi_j)\int_{\mathcal{S}}k\big(\cdot,(s_{\tau,h}^i,a_{\tau,h}^i,s)\big)\mu_{\tau,h}^j(s)\mathrm{d}s\bigg\|_{\mathcal{H}}.
\end{aligned}
$$

To derive a concentration inequality for term (IV), we first construct the minimal $\varepsilon-$cover of $\tilde{\mathcal{W}}$ with respect to $\|\cdot\|_\infty$. The covering number is denoted as $\mathcal{N}_\infty(\varepsilon,\tilde{\mathcal{W}})$. Then for any $W\in\tilde{\mathcal{W}}$, there exists a graphon W_i for $i\in\{1,\cdots,\mathcal{N}_\infty(\varepsilon,\tilde{\mathcal{W}})\}$ such that $\|W-W_i\|_\infty\leq\varepsilon$. Then we have that

$$
\big\|\hat{\omega}_{\tau,h}^i(W)-\bar{\omega}_{\tau,h}^i(W)\big\|_{\mathcal{H}}\leq\big\|\hat{\omega}_{\tau,h}^i(W_i)-\bar{\omega}_{\tau,h}^i(W_i)\big\|_{\mathcal{H}}+2\varepsilon B_k,
$$

where the inequality results from the triangle inequality. In the following, we set $\varepsilon=t/(4B_k)$. Then the concentration inequality for term (IV) can be derived as

$$
\begin{aligned}
&\mathbb{P}\Big(\exists\tilde{W}\in\tilde{\mathcal{W}},i\in[N],\tau\in[L],\big\|\hat{\omega}_{\tau,h}^i(\tilde{W})-\bar{\omega}_{\tau,h}^i(\tilde{W})\big\|_{\mathcal{H}}\geq t) \\
&\quad\leq\mathbb{P}\Big(\exists j\in\big[\mathcal{N}_\infty(\varepsilon,\tilde{\mathcal{W}})\big],i\in[N],\tau\in[L],\big\|\hat{\omega}_{\tau,h}^i(W_j)-\bar{\omega}_{\tau,h}^i(W_j)\big\|_{\mathcal{H}}\geq t-2\varepsilon B_k\Big) \\
&\quad\leq NL\mathcal{N}_\infty(t/(4B_k),\tilde{\mathcal{W}})\max_{j\in[\mathcal{N}_\infty],i\in[N],\tau\in[L]}\mathbb{P}\Big(\big\|\hat{\omega}_{\tau,h}^i(W_j)-\bar{\omega}_{\tau,h}^i(W_j)\big\|_{\mathcal{H}}\geq t/2\Big) \\
&\quad\leq 2NL\mathcal{N}_\infty(t/(4B_k),\tilde{\mathcal{W}})\exp\bigg(-\frac{(N-1)t^2}{32B_k^2}\bigg),
\end{aligned}
$$

where the first inequality results from the construction of the cover, the second inequality results from the union bound, and the last inequality results from Lemma 32 and that

$\|W(\xi_i,\xi_j)k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s^j_{\tau,h})\| \le B_k$ for any $W\in\tilde{\mathcal{W}}$. For $t\ge 4B_k/\sqrt{N}$, we have that

$$\mathbb{P}\Big(\exists \tilde{W}\in\tilde{\mathcal{W}}, i\in[N],\tau\in[L], \big\|\hat{\omega}^i_{\tau,h}(\tilde{W})-\bar{\omega}^i_{\tau,h}(\tilde{W})\big\|_{\mathcal{H}}\ge t\Big) \le 2NL\mathcal{N}_\infty(1/\sqrt{N},\tilde{\mathcal{W}})\exp\bigg(-\frac{(N-1)t^2}{32B_k^2}\bigg).$$

Thus, term (IV) can be bounded as

$$\text{(IV)} = \sup_{W\in\tilde{\mathcal{W}}}\big\|\hat{\omega}^i_{\tau,h}(W)-\bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}} \le \frac{4\sqrt{2}B_k}{\sqrt{N-1}}\log\frac{2NL\mathcal{N}_\infty(1/\sqrt{N},\tilde{\mathcal{W}})}{\delta}, \tag{40}$$

with probability at least $1-\delta$. Substituting inequalities (40) and (39) into inequalities (33) and (34), we have that

$$\sup_{f\in\mathbb{B}(r,\bar{\mathcal{H}}),W\in\tilde{\mathcal{W}}}\bigg|\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1}-f\big(\hat{\omega}^i_{\tau,h}(W)\big)\Big)^2-\Big(s^i_{\tau,h+1}-f\big(\omega^i_{\tau,h}(W)\big)\Big)^2\bigg| \le O\bigg(\frac{(B_S+rB_{\bar{K}})rL_KB_k}{\sqrt{N}}\log\frac{NL\mathcal{N}_\infty(1/\sqrt{N},\tilde{\mathcal{W}})}{\delta}\bigg), \tag{41}$$

with probability at least $1-\delta$.

Step 2: Bound the generalization error of risk.

Considering term (I), for ease of notation, we denote the quadruple $(s^i_{\tau,h},a^i_{\tau,h},\mu^{\mathcal{I}}_{\tau,h},s^i_{\tau,h+1})$ as $e^i_{\tau,h}$. We define the function f_W as

$$f_W(e^i_{\tau,h})=\Big(s^i_{\tau,h+1}-f\big(\omega^i_{\tau,h}(W)\big)\Big)^2-\Big(s^i_{\tau,h+1}-f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2.$$

The correspond function class is defined as $\mathcal{F}_{\tilde{\mathcal{W}}}=\{f_W\,|\,f\in\mathbb{B}(r,\bar{\mathcal{H}}),W\in\tilde{\mathcal{W}}\}$. Then we have that

$$\text{(I)}=\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\big[\hat{f}_{\hat{W}}(e^i_{\tau,h})\big]-2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\hat{f}_{\hat{W}}(e^i_{\tau,h}).$$

Proposition 20 *With Assumption 6, we have that*

$$\begin{aligned}\mathbb{P}\bigg(&\exists f_W\in\mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]-\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}f_W(e^i_{\tau,h})\\ &\ge\varepsilon\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]\bigg)\bigg)\\ &\le 14\mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{\varepsilon\beta}{40(B_S+rB_{\bar{K}})^3B_{\bar{K}}},\mathbb{B}(r,\bar{\mathcal{H}})\bigg)\cdot\mathcal{N}_\infty\bigg(\frac{\varepsilon\beta}{40(B_S+rB_{\bar{K}})^3rL_KB_k},\tilde{\mathcal{W}}\bigg)\\ &\quad\cdot\exp\bigg(-\frac{\varepsilon^2(1-\varepsilon)\alpha NL}{20(B_S+rB_{\bar{K}})^4(1+\varepsilon)}\bigg),\end{aligned}$$

where $\alpha,\beta>0$ and $0<\varepsilon\le 1/2$.

Proof [Proof of Proposition 20] See Appendix Q.1.2. ■

Now consider,

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^i}\big[\hat{f}_{\hat{W}}(e_{\tau,h}^i)\big]-2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\hat{f}_{\hat{W}}(e_{\tau,h}^i)\geq t\bigg)\\
&\quad\leq\mathbb{P}\bigg(\exists f_W\in\mathcal{F}_{\tilde{\mathcal{W}}},\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(e_{\tau,h}^i)\big]-2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}f_W(e_{\tau,h}^i)\geq t\bigg)\\
&\quad=\mathbb{P}\bigg(\exists f_W\in\mathcal{F}_{\tilde{\mathcal{W}}},\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(e_{\tau,h}^i)\big]-\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}f_W(e_{\tau,h}^i)\\
&\qquad\geq\frac{1}{2}\bigg(t+\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(e_{\tau,h}^i)\big]\bigg)\bigg)\\
&\quad\leq 14\mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{t}{160(B_S+rB_{\bar{K}})^3B_{\bar{K}}},\mathbb{B}(r,\bar{\mathcal{H}})\bigg)\cdot\mathcal{N}_\infty\bigg(\frac{t}{160(B_S+rB_{\bar{K}})^3rL_KB_k},\tilde{\mathcal{W}}\bigg)\\
&\qquad\cdot\exp\bigg(-\frac{tNL}{480(B_S+rB_{\bar{K}})^4}\bigg),
\end{aligned}$$

where the last inequality results from Proposition 20. We define that

$$N_{\mathbb{B}_r}=\mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{3}{NL},\mathbb{B}(r,\bar{\mathcal{H}})\bigg),N_{\tilde{\mathcal{W}}}=\mathcal{N}_\infty\bigg(\frac{3}{L_KNL},\tilde{\mathcal{W}}\bigg).$$

For $\delta>0$, we set

$$t=\frac{480(B_S+rB_{\bar{K}})^4}{NL}\log\frac{14N_{\mathbb{B}_r}N_{\tilde{\mathcal{W}}}}{\delta},$$

then we have that

$$\mathbb{P}\bigg(\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^i}\big[\hat{f}_{\hat{W}}(e_{\tau,h}^i)\big]-2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\hat{f}_{\hat{W}}(e_{\tau,h}^i)\geq t\bigg)\leq\delta. \tag{42}$$

Combining inequalities (32), (41), and (42), we have that the following holds with probability at least $1-\delta$

$$\begin{aligned}
&\mathcal{R}_{\bar{\xi}}(\hat{f}_h,\hat{g}_h,\hat{W}_h)-\mathcal{R}_{\bar{\xi}}(f_h^*,g_h^*,W_h^*)\\
&\quad\leq O\bigg(\frac{(B_S+rB_{\bar{K}})^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}+\frac{(B_S+rB_{\bar{K}})rL_KB_k}{\sqrt{N}}\log\frac{NL\mathcal{N}_\infty(1/\sqrt{N},\tilde{\mathcal{W}})}{\delta}\bigg),
\end{aligned}$$

where

$$N_{\tilde{\mathbb{B}}_{\tilde{r}}}=\mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{3}{NL},\mathbb{B}(\tilde{r},\tilde{\mathcal{H}})\bigg).$$

Thus, we conclude the proof of Theorem 5. ■

Appendix H. Proof of Theorem 6

Proof [Proof of Theorem 6] We first decompose the difference between the risk as the sum of the generalization error of risk from position sampling and the difference between the risk given the positions. Our proof involves two steps:

- Bound the generalization error of risk from position sampling.
- Bound the difference between the risk given positions.

$$\begin{aligned}
&\mathcal{R}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}(f_h^*, g_h^*, W_h^*) \\
&\quad = \mathcal{R}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \big(\mathcal{R}(f_h^*, g_h^*, W_h^*) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)\big) \\
&\qquad + \mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*) \\
&\quad \le 2 \sup_{f\in\mathbb{B}(r,\bar{\mathcal{H}}), g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}), W\in\tilde{\mathcal{W}}} \big|\mathcal{R}(f,g,W) - \mathcal{R}_{\bar{\xi}}(f,g,W)\big| + \mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*) \\
&\quad = (\mathrm{IX}) + (\mathrm{X}),
\end{aligned} \tag{43}$$

where (IX) is the generalization error of risk from position sampling, and (X) is the difference between the risk given positions. Similar as the proof of Theorem 5, the terms related to the transition kernels and reward functions in inequality (43) are analogous. In the following, we will only present the proof for the terms related to the transition kernel, and the results for the terms related to the reward functions can be similarly derived.

Step 1: Bound the generalization error of risk from position sampling.

We first define that

$$g_{f,W}(\alpha) = \frac{1}{L}\sum_{\tau=1}^{L} \mathbb{E}_{\rho_{\tau,h}^{\alpha}}\bigg[\Big(s_{\tau,h+1}^{\alpha} - f\big(\omega_{\tau,h}^{\alpha}(W)\big)\Big)^2\bigg].$$

The correspond function class for $g_{f,W}$ is $\mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}} = \{g_{f,W} \mid f \in \mathbb{B}(r,\bar{\mathcal{H}}), W \in \tilde{\mathcal{W}}\}$. Then term in (IX) that is related to the transition kernels can be expressed as

$$2 \sup_{g_{f,W}\in\mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}} \bigg| \int_0^1 g_{f,W}(\alpha)\mathrm{d}\alpha - \frac{1}{N}\sum_{i=1}^{N} g_{f,W}(\xi_i)\bigg|.$$

Let $\delta > 0$, $\mathcal{G}_\delta$ be a minimal L_∞ $\delta-$cover of $\mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}$. Then for any $g_{f,W} \in \mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}$, there exists $\bar{g}_{f,W} \in \mathcal{G}_\delta$ such that $|g_{f,W}(\alpha) - \bar{g}_{f,W}(\alpha)| \le \delta$ for all $\alpha \in \mathcal{I}$. For any $t > 0$, we set $\delta = t/4$. Then we have that

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{g_{f,W}\in\mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}} \bigg| \int_0^1 g_{f,W}(\alpha)\mathrm{d}\alpha - \frac{1}{N}\sum_{i=1}^{N} g_{f,W}(\xi_i)\bigg| \ge t\bigg) \\
&\quad \le \mathcal{N}_\infty\Big(\frac{t}{4}, \mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}\Big) \max_{g_{f,W}\in\mathcal{G}_{\frac{t}{4}}} \mathbb{P}\bigg(\bigg| \int_0^1 g_{f,W}(\alpha)\mathrm{d}\alpha - \frac{1}{N}\sum_{i=1}^{N} g_{f,W}(\xi_i)\bigg| \ge \frac{t}{2}\bigg) \\
&\quad \le 2\mathcal{N}_\infty\Big(\frac{t}{4}, \mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}\Big) \exp\bigg(-\frac{Nt^2}{2(B_S + rB_{\bar{K}})^4}\bigg),
\end{aligned} \tag{44}$$

where the first inequality results from the union bound, and the second inequality results from that $0 \leq g_{f,W}(\alpha) \leq (B_S + rB_{\bar{K}})^2$ and Hoeffding inequality. To upper bound the covering number in the tail probability, we note that

$$\begin{aligned}\big|g_{f,W}(\alpha) - \bar{g}_{f,W}(\alpha)\big| &\leq 2(B_S + rB_{\bar{K}})\frac{1}{L}\sum_{\tau=1}^{L}\mathbb{E}_{\rho^{\alpha}_{\tau,h}}\Big[\Big|f\big(\omega^{\alpha}_{\tau,h}(W)\big) - \bar{f}\big(\omega^{\alpha}_{\tau,h}(\bar{W})\big)\Big|\Big] \\ &\leq 2(B_S + rB_{\bar{K}})\big(B_{\bar{K}}\|f - \bar{f}\|_{\bar{\mathcal{H}}} + rL_K B_k\|W - \bar{W}\|_{\infty}\big),\end{aligned}$$

where the first inequality results from the definition of $g_{f,W}$, and the second inequality results from Lemma 33 and the triangle inequality. This inequality implies that

$$\mathcal{N}_{\infty}\Big(\frac{t}{4}, \mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}\Big) \leq \mathcal{N}_{\bar{\mathcal{H}}}\Big(\frac{t}{16(B_S + rB_{\bar{K}})B_{\bar{K}}}, \mathbb{B}(r, \bar{\mathcal{H}})\Big) \cdot \mathcal{N}_{\infty}\Big(\frac{t}{16(B_S + rB_{\bar{K}})rL_K B_k}, \tilde{\mathcal{W}}\Big).$$

For $1 > \delta > 0$, we take

$$t = \frac{\sqrt{2}(B_S + rB_{\bar{K}})^2}{\sqrt{N}}\log\frac{2\mathcal{N}_{\bar{\mathcal{H}}}\Big(\frac{1}{16\sqrt{N}}, \mathbb{B}(r, \bar{\mathcal{H}})\Big) \cdot \mathcal{N}_{\infty}\Big(\frac{1}{16rL_K B_k\sqrt{N}}, \tilde{\mathcal{W}}\Big)}{\delta}.$$

Then inequality (44) shows that

$$\begin{aligned}&\sup_{g_{f,W}\in\mathcal{G}_{\mathcal{F},\tilde{\mathcal{W}}}}\bigg|\int_0^1 g_{f,W}(\alpha)\mathrm{d}\alpha - \frac{1}{N}\sum_{i=1}^{N} g_{f,W}(\xi_i)\bigg| \\ &= O\bigg(\frac{(B_S + rB_{\bar{K}})^2}{\sqrt{N}}\log\frac{\mathcal{N}_{\bar{\mathcal{H}}}\Big(\frac{1}{16\sqrt{N}}, \mathbb{B}(r, \bar{\mathcal{H}})\Big) \cdot \mathcal{N}_{\infty}\Big(\frac{1}{16rL_K B_k\sqrt{N}}, \tilde{\mathcal{W}}\Big)}{\delta}\bigg),\end{aligned}$$

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

$$\begin{aligned}(\mathrm{IX}) = O\bigg(&\frac{(B_S + rB_{\bar{K}})^2}{\sqrt{N}}\log\frac{\mathcal{N}_{\bar{\mathcal{H}}}\Big(\frac{1}{16\sqrt{N}}, \mathbb{B}(r, \bar{\mathcal{H}})\Big) \cdot \mathcal{N}_{\infty}\Big(\frac{1}{16rL_K B_k\sqrt{N}}, \tilde{\mathcal{W}}\Big)}{\delta} \\ &+ \frac{(B_S + rB_{\tilde{K}})^2}{\sqrt{N}}\log\frac{\mathcal{N}_{\tilde{\mathcal{H}}}\Big(\frac{1}{16\sqrt{N}}, \mathbb{B}(\tilde{r}, \tilde{\mathcal{H}})\Big) \cdot \mathcal{N}_{\infty}\Big(\frac{1}{16rL_K B_k\sqrt{N}}, \tilde{\mathcal{W}}\Big)}{\delta}\bigg) \qquad (45)\end{aligned}$$

Step 2: Bound the difference between the risk given positions.

We adopt the similar procedures as the proof of Theorem 5. From inequalities (32), (33),and (34), we define that

$$\begin{aligned}(\mathrm{XI}) = \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\bigg[&\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h)\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\bigg] \\ &- 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h)\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\end{aligned}$$

$$\text{(XII)} = 4(B_S + rB_{\bar{K}})rL_K \sup_{W\in\tilde{\mathcal{W}}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\big\|\hat{\omega}^i_{\tau,h}(W) - \bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}}$$
$$\text{(XIII)} = 4(B_S + rB_{\bar{K}})rL_K \sup_{W\in\tilde{\mathcal{W}}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\big\|\bar{\omega}^i_{\tau,h}(W) - \omega^i_{\tau,h}(W)\big\|_{\mathcal{H}}.$$

For term (XIII), we adopt a different method with the proof of Theorem 5. Let $\varepsilon > 0$, $\tilde{\mathcal{W}}_\varepsilon$ be a L_∞ $\varepsilon-$cover of $\tilde{\mathcal{W}}$. Then for any $W \in \tilde{\mathcal{W}}$, there exists $\bar{W} \in \tilde{\mathcal{W}}_\varepsilon$ such that $\|\bar{W} - W\|_\infty \le \varepsilon$. Then we have

$$\begin{aligned}\|\omega^i_{\tau,h}(W) - \omega^i_{\tau,h}(\bar{W})\|_{\mathcal{H}} &= \bigg\|\int_0^1\int_{\mathcal{S}}\big(W(\xi_i,\beta) - \bar{W}(\xi_i,\beta)\big)k\big(\cdot,(s^i_{\tau,h},a^i_{\tau,h},s)\big)\mu^\beta_{\tau,h}(s)\mathrm{d}s\mathrm{d}\beta\bigg\|_{\mathcal{H}}\\ &\le \varepsilon B_k,\end{aligned}$$

where the inequality results from the triangle inequality. Similarly, we have that $\|\bar{\omega}^i_{\tau,h}(W) - \bar{\omega}^i_{\tau,h}(\bar{W})\|_{\mathcal{H}} \le \varepsilon B_k$. For any $t > 0$, we will set $\varepsilon = t/(4B_k)$. Then the tail probability for (XIII) can be bounded as

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{W\in\tilde{\mathcal{W}}}\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\big\|\bar{\omega}^i_{\tau,h}(W) - \omega^i_{\tau,h}(W)\big\|_{\mathcal{H}} \ge t\bigg)\\ &\quad\le \mathbb{P}\bigg(\exists W\in\tilde{\mathcal{W}}, \tau\in[L], i\in[N], \big\|\bar{\omega}^i_{\tau,h}(W) - \omega^i_{\tau,h}(W)\big\|_{\mathcal{H}} \ge t\bigg)\\ &\quad\le NL\mathcal{N}_\infty\bigg(\frac{t}{4B_k},\tilde{\mathcal{W}}\bigg)\max_{W\in\tilde{\mathcal{W}}_{\frac{t}{4B_k}},\tau\in[L],i\in[N]}\mathbb{P}\bigg(\big\|\bar{\omega}^i_{\tau,h}(W) - \omega^i_{\tau,h}(W)\big\|_{\mathcal{H}} \ge \frac{t}{2}\bigg)\\ &\quad\le 2NL\mathcal{N}_\infty\bigg(\frac{t}{4B_k},\tilde{\mathcal{W}}\bigg)\exp\bigg(-\frac{(N-1)t^2}{8B_k^2}\bigg),\end{aligned}$$

where the second inequality results from the union bound, and the last inequality resutls from Lemma 32. For any $0 < \delta < 1$, we set

$$t = \frac{2\sqrt{2}B_k}{\sqrt{N-1}}\log\frac{2NL\mathcal{N}_\infty\Big(\frac{1}{\sqrt{N}},\tilde{\mathcal{W}}\Big)}{\delta}.$$

Then we have that

$$\text{(XIII)} \le O\Bigg(\frac{(B_S + rB_{\bar{K}})rL_KB_k}{\sqrt{N}}\log\frac{NL\mathcal{N}_\infty\Big(\frac{1}{\sqrt{N}},\tilde{\mathcal{W}}\Big)}{\delta}\Bigg) \tag{46}$$

with probability at least $1 - \delta$.

For term (XII), we follow the proof of Theorem 5 and condition on the values of $\bar{\xi}$ to bound the tail probability. We have that

$$\mathbb{P}\bigg(\sup_{W\in\tilde{\mathcal{W}}}\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\big\|\hat{\omega}^i_{\tau,h}(W) - \bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}} \ge t\bigg)$$

$$= \mathbb{E}_{\bar{\xi}}\bigg[\mathbb{P}\bigg(\sup_{W\in\tilde{\mathcal{W}}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\big\|\hat{\omega}^i_{\tau,h}(W)-\bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}} \geq t \,\bigg|\, \bar{\xi}\bigg)\bigg]$$
$$\leq 2NL\mathcal{N}_{\infty}(1/\sqrt{N},\tilde{\mathcal{W}})\exp\bigg(-\frac{(N-1)t^2}{32B_k^2}\bigg),$$

where we condition on the values of $\bar{\xi}$ in the first equality, and the inequality results from inequality (40). Thus, we have that

$$\text{(XII)} \leq O\bigg(\frac{(B_S+rB_{\bar{K}})rL_K B_k}{\sqrt{N}}\log\frac{NL\mathcal{N}_{\infty}(1/\sqrt{N},\tilde{\mathcal{W}})}{\delta}\bigg) \tag{47}$$

with probability at least $1-\delta$.

For term (XI), we just adopt the same conditional probability trick as shown in the bound of (XII). From inequality (42), we have that

$$\text{(XI)} \leq O\bigg(\frac{(B_S+rB_{\bar{K}})^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathcal{W}}}}{\delta}\bigg) \tag{48}$$

with probability at least $1-\delta$.

Combining the inequalities (43), (45), (46), (47), and (48), we have that

$$\begin{aligned}
&\mathcal{R}(\hat{f}_h,\hat{g}_h,\hat{W}_h)-\mathcal{R}(f_h^*,g_h^*,W_h^*)\\
&\quad\leq O\bigg(\frac{\big(B_S+r\max\{B_{\bar{K}},B_{\tilde{K}}\}\big)^2}{\sqrt{N}}\log\frac{\tilde{N}_{\mathbb{B}_r}\tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}}\tilde{N}_{\tilde{\mathcal{W}}}}{\delta}\\
&\qquad+\frac{\big(B_S+r\max\{B_{\bar{K}},B_{\tilde{K}}\}\big)r\max\{L_K,L_{\tilde{K}}B_{\bar{K}}\}}{\sqrt{N}}\log\frac{NL\mathcal{N}_{\infty}\Big(\frac{1}{\sqrt{N}},\tilde{\mathcal{W}}\Big)}{\delta}\\
&\qquad+\frac{(B_S+r\max\{B_{\bar{K}},B_{\tilde{K}})^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}\bigg).
\end{aligned}$$

Thus, we conclude the proof of Theorem 6. ∎

Appendix I. Proof of Theorem 7

Proof [Proof of Theorem 7] We first decompose the difference between the permutation-invariant risk as the sum of the generalization error of risk, the Estimation Error of Mean-embedding, and the empirical risk difference. Given the fact that the empirical risk difference is equal and less to zero, Our proof involves two steps:

- Bound the estimation error of mean-embedding.
- Bound the generalization error of the risk.

Consider,

$$\bar{\mathcal{R}}_{\bar{\xi}}(\hat{f}_h,\hat{g}_h,\hat{W}_h)-\bar{\mathcal{R}}_{\bar{\xi}}(f_h^*,g_h^*,W_h^*)$$

$$
\begin{aligned}
&= \inf_{\phi\in\mathcal{B}_{[0,1]}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\Big[\Big(s^i_{\tau,h+1}-\hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2+\Big(r^i_{\tau,h}-\hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2 \\
&\quad -\Big(s^i_{\tau,h+1}-f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2-\Big(r^i_{\tau,h}-g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\Big] \\
&\le \inf_{\phi\in\mathcal{C}^N_{[0,1]}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\Big[\Big(s^i_{\tau,h+1}-\hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2+\Big(r^i_{\tau,h}-\hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2 \\
&\quad -\Big(s^i_{\tau,h+1}-f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2-\Big(r^i_{\tau,h}-g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\Big] \\
&= \text{Generalization Error of Risk} + \text{Estimation Error of Mean-embedding} \\
&\qquad + \text{Empirical Risk Difference},
\end{aligned}
$$

where each term is defined as

$$
\begin{aligned}
&\text{Generalization Error of Risk} \\
&= \inf_{\phi\in\mathcal{C}^N_{[0,1]}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\Big[\Big(s^i_{\tau,h+1}-\hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2+\Big(r^i_{\tau,h}-\hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2 \\
&\quad -\Big(s^i_{\tau,h+1}-f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2-\Big(r^i_{\tau,h}-g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\Big] \\
&\quad -2\inf_{\phi\in\mathcal{C}^N_{[0,1]}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big[\Big(s^i_{\tau,h+1}-\hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2+\Big(r^i_{\tau,h}-\hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2 \\
&\quad -\Big(s^i_{\tau,h+1}-f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2-\Big(r^i_{\tau,h}-g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\Big].
\end{aligned}
$$

This generalization error of risk represents the error due to the fact that we optimize over the empirical estimation of the risk not the population risk.

$$
\begin{aligned}
&\text{Estimation Error of Mean-embedding} \\
&= 2\inf_{\phi\in\mathcal{C}^N_{[0,1]}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1}-\hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2+\Big(r^i_{\tau,h}-\hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2 \\
&\quad -2\inf_{\phi\in\mathcal{C}^N_{[0,1]}} \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1}-\hat{f}_h\big(\hat{\bar{\omega}}^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2+\Big(r^i_{\tau,h}-\hat{g}_h\big(\hat{\bar{\omega}}^i_{\tau,h}(\hat{W}^{\phi}_h)\big)\Big)^2 \\
&\quad +2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1}-f^*_h\big(\hat{\bar{\omega}}^i_{\tau,h}(W^{*,\phi^*}_h)\big)\Big)^2+\Big(r^i_{\tau,h}-g^*_h\big(\hat{\bar{\omega}}^i_{\tau,h}(W^{*,\phi^*}_h)\big)\Big)^2 \\
&\quad -2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1}-f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2+\Big(r^i_{\tau,h}-g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2.
\end{aligned}
$$

Estimation error of mean-embedding represents the error due to the fact that we cannot observe the value of $\hat{\omega}^i_{\tau,h}(\hat{W}_h)$. Instead, we can only estimate the value of it through the states of sampled agents.

Empirical Risk Difference

$$= 2 \inf_{\phi \in \mathcal{C}_{[0,1]}^N} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - \hat{f}_h\big(\hat{\hat{\omega}}_{\tau,h}^i(\hat{W}_h^{\phi})\big)\Big)^2 + \Big(r_{\tau,h}^i - \hat{g}_h\big(\hat{\hat{\omega}}_{\tau,h}^i(\hat{W}_h^{\phi})\big)\Big)^2$$
$$- 2 \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - f_h^*\big(\hat{\hat{\omega}}_{\tau,h}^i(W_h^{*,\phi^*})\big)\Big)^2 + \Big(r_{\tau,h}^i - g_h^*\big(\hat{\hat{\omega}}_{\tau,h}^i(W_h^{*,\phi^*})\big)\Big)^2,$$

where $\phi^* \in \mathcal{C}_{[0,1]}^N$ is a permutation of $((i-1)/N, i/N]$ for $i \in [N]$ such that $\phi^*(i/N) = \xi_i$. From the estimation procedure of Algorithm (10), we have that

$$\text{Empirical Risk Difference} \leq 0.$$

Thus, we have that

$$\bar{\mathcal{R}}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \bar{\mathcal{R}}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)$$
$$\leq \text{Generalization Error of Risk} + \text{Estimation Error of Mean-embedding}.$$

Step 1: Bound the Estimation Error of Mean-embedding.
From the definition of the generalization error, we bound two terms separately

$$\text{Estimation Error of Mean-embedding}$$
$$\leq 2 \sup_{f,g,W} \bigg| \inf_{\phi \in \mathcal{C}_{[0,1]}^N} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - f\big(\hat{\hat{\omega}}_{\tau,h}^i(W^{\phi})\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\hat{\hat{\omega}}_{\tau,h}^i(W^{\phi})\big)\Big)^2$$
$$- \inf_{\phi \in \mathcal{C}_{[0,1]}^N} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - f\big(\omega_{\tau,h}^i(W^{\phi})\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\omega_{\tau,h}^i(W^{\phi})\big)\Big)^2 \bigg|$$
$$+ 2 \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - f_h^*\big(\hat{\hat{\omega}}_{\tau,h}^i(W_h^{*,\phi^*})\big)\Big)^2 + \Big(r_{\tau,h}^i - g_h^*\big(\hat{\hat{\omega}}_{\tau,h}^i(W_h^{*,\phi^*})\big)\Big)^2$$
$$- 2 \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - f_h^*\big(\omega_{\tau,h}^i(W_h^*)\big)\Big)^2 + \Big(r_{\tau,h}^i - g_h^*\big(\omega_{\tau,h}^i(W_h^*)\big)\Big)^2$$
$$= (\text{XIV}) + (\text{XV}) \tag{49}$$

We first denote the composition of two measure-preserving bijections ϕ and ψ as $\phi \circ \psi$. When appiled to a graphon W, the composition of bijections maps the values of the graphon as

$$W^{\phi \circ \psi}(x, y) = W\Big(\phi\big(\psi(x)\big), \phi\big(\psi(y)\big)\Big).$$

Then we bound the term (XIV) as

$$\sup_{f,g,W} \bigg| \inf_{\phi \in \mathcal{C}_{[0,1]}^N} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - f\big(\hat{\hat{\omega}}_{\tau,h}^i(W^{\phi \circ \phi^*})\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\hat{\hat{\omega}}_{\tau,h}^i(W^{\phi \circ \phi^*})\big)\Big)^2$$
$$- \inf_{\phi \in \mathcal{C}_{[0,1]}^N} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^i - f\big(\omega_{\tau,h}^i(W^{\phi})\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\omega_{\tau,h}^i(W^{\phi})\big)\Big)^2 \bigg|$$

$$
\begin{aligned}
&= \sup_{f,W,\phi} \bigg| \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^{i} - f\big(\hat{\bar{\omega}}_{\tau,h}^{i}(W^{\phi\circ\phi^*})\big)\Big)^2 + \Big(r_{\tau,h}^{i} - g\big(\hat{\bar{\omega}}_{\tau,h}^{i}(W^{\phi\circ\phi^*})\big)\Big)^2 \\
&\qquad - \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s_{\tau,h+1}^{i} - f\big(\omega_{\tau,h}^{i}(W^{\phi})\big)\Big)^2 + \Big(r_{\tau,h}^{i} - g\big(\omega_{\tau,h}^{i}(W^{\phi})\big)\Big)^2 \bigg| \\
&\leq 4(B_S + \bar{r}B_K)\bar{r}L_K \sup_{W\in\tilde{\mathcal{W}},\phi\in\mathcal{C}_{[0,1]}^N} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \big\| \hat{\bar{\omega}}_{\tau,h}^{i}(W^{\phi\circ\phi^*}) - \omega_{\tau,h}^{i}(W^{\phi}) \big\|_{\mathcal{H}}, \qquad (50)
\end{aligned}
$$

where the equality results from the fact that ϕ^* is a measure-preserving bijection, and the inequality results from the same arguments in inequality (33).

We decompose the error as

$$
\begin{aligned}
&\sup_{W,\phi} \big\| \hat{\bar{\omega}}_{\tau,h}^{i}(W^{\phi\circ\phi^*}) - \omega_{\tau,h}^{i}(W^{\phi}) \big\|_{\mathcal{H}} \\
&\quad \leq \sup_{W,\phi} \big\| \bar{\omega}_{\tau,h}^{i}(W^{\phi}) - \omega_{\tau,h}^{i}(W^{\phi}) \big\|_{\mathcal{H}} + \sup_{W,\phi} \big\| \hat{\bar{\omega}}_{\tau,h}^{i}(W^{\phi\circ\phi^*}) - \bar{\omega}_{\tau,h}^{i}(W^{\phi}) \big\|_{\mathcal{H}}
\end{aligned}
$$

where

$$
\begin{aligned}
\omega_{\tau,h}^{i}(W^{\phi}) &= \int_0^1 \int_{\mathcal{S}} W\big(\phi(\xi_i),\phi(\beta)\big) k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s)\big) \mu_{\tau,h}^{\beta}(s)\mathrm{d}s\mathrm{d}\beta, \\
\bar{\omega}_{\tau,h}^{i}(W^{\phi}) &= \frac{1}{N-1} \sum_{j\neq i} W\big(\phi(\xi_i),\phi(\xi_j)\big) \int_{\mathcal{S}} k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s)\big) \mu_{\tau,h}^{j}(s)\mathrm{d}s, \\
\hat{\bar{\omega}}_{\tau,h}^{i}(W^{\phi\circ\phi^*}) &= \frac{1}{(N-1)L} \sum_{j\neq i} \sum_{\tau'=1}^{L} W\big(\phi(\xi_i),\phi(\xi_j)\big) k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s_{\tau',h}^{j})\big).
\end{aligned}
$$

For term $\sup_{W,\phi} \big\| \bar{\omega}_{\tau,h}^{i}(W^{\phi}) - \omega_{\tau,h}^{i}(W^{\phi}) \big\|_{\mathcal{H}}$, we define the interval $\mathcal{I}_i = ((i-1)/N, i/N]$ for $i \in [N]$. Then we have that

$$
\begin{aligned}
&\big\| \bar{\omega}_{\tau,h}^{i}(W^{\phi}) - \omega_{\tau,h}^{i}(W^{\phi}) \big\|_{\mathcal{H}} \\
&\quad \leq \frac{2}{N}B_k + \sum_{j\neq i} \bigg\| \int_{\xi_j-\frac{1}{N}}^{\xi_j} \int_{\mathcal{S}} W\big(\phi(\xi_i),\phi(\beta)\big) k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s)\big) \mu_{\tau,h}^{\beta}(s)\mathrm{d}s\mathrm{d}\beta \\
&\qquad\qquad - \frac{1}{N} \sum_{j\neq i} W\big(\phi(\xi_i),\phi(\xi_j)\big) \int_{\mathcal{S}} k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s)\big) \mu_{\tau,h}^{j}(s)\mathrm{d}s \bigg\|_{\mathcal{H}},
\end{aligned}
$$

where the inequality results from the triangle inequality. For each term in the sum, we bound it as

$$
\begin{aligned}
&\bigg\| \int_{\xi_j-\frac{1}{N}}^{\xi_j} \int_{\mathcal{S}} W\big(\phi(\xi_i),\phi(\beta)\big) k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s)\big) \mu_{\tau,h}^{\beta}(s)\mathrm{d}s\mathrm{d}\beta \\
&\quad - \frac{1}{N} \sum_{j\neq i} W\big(\phi(\xi_i),\phi(\xi_j)\big) \int_{\mathcal{S}} k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s)\big) \mu_{\tau,h}^{j}(s)\mathrm{d}s \bigg\|_{\mathcal{H}} \\
&\leq \bigg\| \int_{\xi_j-\frac{1}{N}}^{\xi_j} \int_{\mathcal{S}} W\big(\phi(\xi_i),\phi(\beta)\big) k\big(\cdot,(s_{\tau,h}^{i},a_{\tau,h}^{i},s)\big) \big(\mu_{\tau,h}^{\beta}(s) - \mu_{\tau,h}^{j}(s)\big)\mathrm{d}s\mathrm{d}\beta \bigg\|_{\mathcal{H}}
\end{aligned}
$$

$$
\begin{aligned}
&+\bigg\|\int_{\xi_j-\frac{1}{N}}^{\xi_j}\int_{\mathcal{S}}\Big(W\big(\phi(\xi_i),\phi(\beta)\big)-W\big(\phi(\xi_i),\phi(\xi_j)\big)\Big)k\big(\cdot,(s_{\tau,h}^i,a_{\tau,h}^i,s)\big)\mu_{\tau,h}^j(s)\mathrm{d}s\mathrm{d}\beta\bigg\|_{\mathcal{H}}\\
&=O\bigg(\frac{L_{\mathcal{W}}B_k}{N^2}\bigg),
\end{aligned}
$$

where the first inequality results from the triangle inequality, and the second inequality results from the same argument in inequality (38) and the fact that β and ξ_j are always in the same interval for any $\phi\in\mathcal{C}_{[0,1]}^N$. Thus, we have that

$$
\sup_{W,\phi}\big\|\bar{\omega}_{\tau,h}^i(W^\phi)-\omega_{\tau,h}^i(W^\phi)\big\|_{\mathcal{H}}=O\bigg(\frac{L_{\mathcal{W}}B_k}{N}\bigg). \tag{51}
$$

For $\sup_{W,\phi}\big\|\hat{\bar{\omega}}_{\tau,h}^i(W^{\phi\circ\phi^*})-\bar{\omega}_{\tau,h}^i(W^\phi)\big\|_{\mathcal{H}}$, we adopt the similar procedure in the proof of inequality (40).

$$
\begin{aligned}
&\mathbb{P}\Big(\sup_{W,\phi}\big\|\hat{\bar{\omega}}_{\tau,h}^i(W^{\phi\circ\phi^*})-\bar{\omega}_{\tau,h}^i(W^\phi)\big\|_{\mathcal{H}}\geq t\Big)\\
&\quad\leq N!NL\mathcal{N}_\infty(t/(4B_k),\tilde{\mathcal{W}})\max_{j\in[\mathcal{N}_\infty],i\in[N],\tau\in[L]}\mathbb{P}\Big(\big\|\hat{\bar{\omega}}_{\tau,h}^i(W_j^{\phi\circ\phi^*})-\bar{\omega}_{\tau,h}^i(W_j^\phi)\big\|_{\mathcal{H}}\geq t/2\Big)\\
&\quad\leq 2N!NL\mathcal{N}_\infty(t/(4B_k),\tilde{\mathcal{W}})\exp\bigg(-\frac{NLt^2}{16B_k^2}\bigg),
\end{aligned}
$$

where the first inequality results from the proof of inequality (40), and the last inequality results from Lemma 32. Thus, we have that with probability at least $1-\delta$

$$
\sup_{W,\phi}\big\|\hat{\bar{\omega}}_{\tau,h}^i(W^{\phi\circ\phi^*})-\bar{\omega}_{\tau,h}^i(W^\phi)\big\|_{\mathcal{H}}=O\bigg(B_k\sqrt{\frac{N}{L}}\log\frac{NL\mathcal{N}_\infty(\sqrt{N/L},\tilde{\mathcal{W}})}{\delta}\bigg). \tag{52}
$$

Combining inequalities (50), (51) and (52), we have that

$$
(\mathrm{XIV})=O\bigg(\frac{L_{\mathcal{W}}B_k\bar{r}L_K(B_S+\bar{r}B_K)}{N}+(B_S+\bar{r}B_K)\bar{r}L_KB_k\sqrt{\frac{N}{L}}\log\frac{NL\mathcal{N}_\infty(\sqrt{N/L},\tilde{\mathcal{W}})}{\delta}\bigg). \tag{53}
$$

Following the similar arguments, we can derive that

$$
(\mathrm{XV})=O\bigg(\frac{B_k\bar{r}L_K(B_S+\bar{r}B_K)}{N}+(B_S+\bar{r}B_K)\bar{r}L_KB_k\frac{1}{\sqrt{NL}}\log\frac{NL\mathcal{N}_\infty(\sqrt{N/L},\tilde{\mathcal{W}})}{\delta}\bigg).
$$

Step 2: Bound the generalization error of risk

We follow the similar procedures in Step 2 of the proof of Theorem 5. We denote the quadruple $(s_{\tau,h}^i,a_{\tau,h}^i,\mu_{\tau,h}^{\mathcal{I}},s_{\tau,h+1}^i)$ as $e_{\tau,h}^i$. We define the function f_W as

$$
f(e_{\tau,h}^i,W,\phi)=\Big(s_{\tau,h+1}^i-f\big(\omega_{\tau,h}^i(W^\phi)\big)\Big)^2-\Big(s_{\tau,h+1}^i-f_h^*\big(\omega_{\tau,h}^i(W_h^*)\big)\Big)^2.
$$

Then we have that

$$
\mathbb{P}\bigg(\inf_{\phi\in\mathcal{C}_{[0,1]}^N}\frac{1}{NL}\sum_{\tau=1}^L\sum_{i=1}^N\mathbb{E}_{\rho_{\tau,h}^i}\big[\hat{f}_h(e_{\tau,h}^i,\hat{W}_h,\phi)\big]-2\inf_{\phi\in\mathcal{C}_{[0,1]}^N}\frac{1}{NL}\sum_{\tau=1}^L\hat{f}_h(e_{\tau,h}^i,\hat{W}_h,\phi)\geq t\bigg)
$$

$$\le \mathbb{P}\bigg(\exists f{\in}\mathbb{B}(r,\bar{\mathcal{H}}), W{\in}\tilde{\mathcal{W}}, \max_{\phi\in\mathcal{C}^N_{[0,1]}}\Big[\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\big[f(e^i_{\tau,h},W,\phi)\big]-\frac{2}{NL}\sum_{\tau=1}^{L}f(e^i_{\tau,h},W,\phi)\Big]\ge t\bigg)$$
$$\le N!\max_{\phi\in\mathcal{C}^N_{[0,1]}}\mathbb{P}\bigg(\exists f{\in}\mathbb{B}(r,\bar{\mathcal{H}}), W{\in}\tilde{\mathcal{W}}, \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\big[f(e^i_{\tau,h},W,\phi)\big]-\frac{2}{NL}\sum_{\tau=1}^{L}f(e^i_{\tau,h},W,\phi)\ge t\bigg)$$
$$\le 14N!\mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{t}{160(B_S+rB_{\bar{K}})^3B_{\bar{K}}},\mathbb{B}(r,\bar{\mathcal{H}})\bigg)\cdot\mathcal{N}_\infty\bigg(\frac{t}{160(B_S+rB_{\bar{K}})^3rL_KB_k},\tilde{\mathcal{W}}\bigg),$$

where the second inequality results from the union bound and the fact that $\min_x f(x) - \min_x g(x) \le \max_x f(x) - g(x)$, and the final inequality results from Proposition 20. Thus, we have that with probability at least $1-\delta$

$$\inf_{\phi\in\mathcal{C}^N_{[0,1]}}\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\big[\hat{f}_h(e^i_{\tau,h},\hat{W}_h,\phi)\big]-2\inf_{\phi\in\mathcal{C}^N_{[0,1]}}\frac{1}{NL}\sum_{\tau=1}^{L}\hat{f}_h(e^i_{\tau,h},\hat{W}_h,\phi)$$
$$=O\bigg(\frac{(B_S+rB_{\bar{K}})^4}{L}\log\frac{N\tilde{N}_{\mathbb{B}_r}\tilde{N}_\infty}{\delta}\bigg), \tag{54}$$

where

$$\tilde{N}_{\mathbb{B}_r}=\mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{3}{L},\mathbb{B}(r,\bar{\mathcal{H}})\bigg),\tilde{N}_{\tilde{\mathcal{W}}}=\mathcal{N}_\infty\bigg(\frac{3}{L_KL},\tilde{\mathcal{W}}\bigg).$$

Combining inequalities (54) and (53), we have that

$$\bar{\mathcal{R}}_{\bar{\xi}}(\hat{f}_h,\hat{g}_h,\hat{W}_h)-\bar{\mathcal{R}}_{\bar{\xi}}(f^*_h,g^*_h,W^*_h)$$
$$=O\bigg(\frac{L_{\mathcal{W}}B_k\bar{r}L_K(B_S+\bar{r}B_K)}{N}+(B_S+\bar{r}B_K)\bar{r}L_KB_K\sqrt{\frac{N}{L}}\log\frac{NL\mathcal{N}_\infty(\sqrt{N/L},\tilde{\mathcal{W}})}{\delta}$$
$$+\frac{(B_S+\bar{r}B_K)^4}{L}\log\frac{N\tilde{N}_{\mathbb{B}_r}\tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}}\tilde{N}_\infty}{\delta}\bigg),$$

where

$$\tilde{N}_{\tilde{\mathbb{B}}_{\tilde{r}}}=\mathcal{N}_{\tilde{\mathcal{H}}}\bigg(\frac{3}{L},\mathbb{B}(\tilde{r},\tilde{\mathcal{H}})\bigg).$$

Thus, we conclude the proof of Theorem 7.

■

Appendix J. Proof of Corollary 12

Proof [Proof of Corollary 12] The proof of Corollary 12 follows the same procedures as the proof of Theorem 5. The only difference is that inequality (40) in the proof of Theorem 5 is replaced by

$$(\mathrm{IV})=\sup_{W\in\tilde{\mathcal{W}}}\big\|\hat{\omega}^i_{\tau,h}(W)-\bar{\omega}^i_{\tau,h}(W)\big\|_{\mathcal{H}}\le\frac{4\sqrt{2}B_k}{\sqrt{(N-1)L}}\log\frac{2NL\mathcal{N}_\infty(1/\sqrt{NL},\tilde{\mathcal{W}})}{\delta}.$$

■

Appendix K. Proof of Corollary 9

Proof [Proof of Corollary 9] Our proof mainly involves four steps

- Derive the performance guarantee of Algorithm (12).
- Generalize the performance guarantee from $\{\xi_i\}_i^N$ to $[0,1]$ by lipschitzness.
- Bound the estimation error of distribution flow and action-value function estimate.
- Conclude the final result.

Step 1: Derive the performance guarantee of Algorithm (12).

We first derive the performance guarantee of Algorithm (12) when we we sample agents with known grid positions. In such setting, we implement $\pi^{\mathcal{I}}$ for L times on the MDP induced by $\mu^{\mathcal{I}}$ to collect the dataset $\mathcal{D}_\tau = \{(s_{\tau,h}^{[N]}, a_{\tau,h}^{[N]}, r_{\tau,h}^{[N]}, s_{\tau,h+1}^{[N]})\}_{h=1}^H$ for $\tau \in [L]$. We define $\mu^{+,\mathcal{I}} = \Gamma_3(\pi^{\mathcal{I}}, \mu^{\mathcal{I}}, W^*)$ as the distribution flow of implementing $\pi^{\mathcal{I}}$ on the MDP induced by $\mu^{\mathcal{I}}$. Then the joint distribution of $(s_{\tau,h}^i, a_{\tau,h}^i, r_{\tau,h}^i, s_{\tau,h+1}^i)_{i=1}^N$ is $\prod_{i=1}^N \rho_{\tau,h}^{+,i}$, where $\rho_{\tau,h}^{+,i} = \mu_{\tau,h}^{+,i} \times \pi_{\tau,h}^i \times \delta_{r_h^*} \times P_h^*$. With a little abuse of notation, we define the risk of (f,g,W) given $\bar{\xi}$ as

$$\mathcal{R}_{\bar{\xi}}(f,g,W) = \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho_{\tau,h}^{+,i}}\Big[\Big(s_{\tau,h+1}^i - f\big(\omega_{\tau,h}^i(W)\big)\Big)^2 + \Big(r_{\tau,h}^i - g\big(\omega_{\tau,h}^i(W)\big)\Big)^2\Big]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\rho_h^{+,i}}\Big[\Big(s_{h+1}^i - f\big(\omega_h^i(W)\big)\Big)^2 + \Big(r_{h+1}^i - g\big(\omega_h^i(W)\big)\Big)^2\Big],$$

where the second equality results from that we implement the same policy for L times. The difference between this definition and Eqn. (7) is that we take expectation with respect to $\rho_{\tau,h}^{+,i}$ instead of $\rho_{\tau,h}^i$. The reason is that in the setting where we specify Eqn. (7), the MDP is induced by the distribution flow of the policy itself, not by a pre-specified distribution flow. We state the performance guarantee as

Corollary 21 *Under Assumptions 5, 6, 7, and 1, if $\xi_i = i/N$ for $i \in [N]$, then the risk of estimate derived in Algorithm (12) can be bounded as*

$$\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)$$
$$\le O\bigg(\frac{(B_S + \bar{r}B_K)^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}\bigg)$$

with probability at least $1-\delta$, where $N_{\mathbb{B}_r}$, $N_{\tilde{\mathbb{B}}_{\tilde{r}}}$, and $N_{\tilde{\mathcal{W}}}$ are defined in Theorem 5.

Proof [Proof of Corollary 21] See Appendix Q.3.1. ■

Step 2: Generalize the performance guarantee from $\{\xi_i\}_{i=1}^N$ to $[0,1]$ by lipschitzness.

Intuitively, when the implemented policy is lipschitz, we can generalize the performance guarantee of $\mathcal{R}_{\bar{\xi}}(f,g,W)$ to that of $\mathcal{R}(f,g,W)$. Here we consider the case where the MDP is induced by the distribution flow of the policy itself, i.e., the case specified in Section 5. The results for the case where the MDP is induced by a pre-specified distribution flow can be similarly derived. We note that

$$\begin{aligned}
&\mathcal{R}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}(f_h^*, g_h^*, W_h^*) \\
&\quad = \mathcal{R}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \big(\mathcal{R}(f_h^*, g_h^*, W_h^*) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)\big) \\
&\qquad + \mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*) \\
&\quad \leq 2 \sup_{f\in\mathbb{B}(r,\bar{\mathcal{H}}), g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}), W\in\tilde{\mathcal{W}}} \big|\mathcal{R}(f,g,W) - \mathcal{R}_{\bar{\xi}}(f,g,W)\big| + \mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*).
\end{aligned} \tag{55}$$

Then we attempt to bound the first term of the right-hand side of inequality (55). For any two positions $\alpha, \beta \in \mathcal{I}$ and $f \in \mathbb{B}(r, \bar{\mathcal{H}})$, we have

$$\begin{aligned}
&\bigg|\mathbb{E}_{\rho_{\tau,h}^\alpha}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big] - \mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\beta(W)\big)\Big)^2\Big]\bigg| \\
&\quad \leq \bigg|\mathbb{E}_{\rho_{\tau,h}^\alpha}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big] - \mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big]\bigg| \\
&\qquad + \bigg|\mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big] - \mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\beta(W)\big)\Big)^2\Big]\bigg|,
\end{aligned} \tag{56}$$

where the inequality results from the triangle inequality. For the first term in the right-hand side of inequality (56), we have that

$$\begin{aligned}
&\bigg|\mathbb{E}_{\rho_{\tau,h}^\alpha}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big] - \mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big]\bigg| \\
&\quad \leq (B_S + rB_{\bar{K}})^2\Big[\|\mu_{\tau,h}^\alpha - \mu_{\tau,h}^\beta\|_1 + \mathbb{E}_{\mu_{\tau,h}^\alpha}\big[\|\pi_{\tau,h}^\alpha(\cdot\,|\,s) - \pi_{\tau,h}^\beta(\cdot\,|\,s)\|_1\big] \\
&\qquad + L_P\big\|z_h^\alpha(\mu_{\tau,h}^{\mathcal{I}}, W_h^*) - z_h^\beta(\mu_{\tau,h}^{\mathcal{I}}, W_h^*)\big\|_1\Big] \\
&\quad \leq C(B_S + rB_{\bar{K}})^2 \cdot |\alpha - \beta|,
\end{aligned}$$

where $C > 0$ is a constant, the first inequality results from the definition of $\rho_{\tau,h}^{\mathcal{I}}$, and the last inequality adopts Proposition 19 and Assumption 5 to bound these three terms. The second term in the right-hand side of inequality (56) can be bounded as

$$\begin{aligned}
&\bigg|\mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big] - \mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\beta(W)\big)\Big)^2\Big]\bigg| \\
&\quad \leq 2(B_S + rB_{\bar{K}})rL_K L_{\mathcal{W}} B_k|\alpha - \beta|,
\end{aligned}$$

where the inequality results from Lemma 33 and Assumption 5. Thus, we conclude that

$$\begin{aligned}
&\bigg|\mathbb{E}_{\rho_{\tau,h}^\alpha}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\alpha(W)\big)\Big)^2\Big] - \mathbb{E}_{\rho_{\tau,h}^\beta}\Big[\Big(s_{\tau,h+1} - f\big(\omega_{\tau,h}^\beta(W)\big)\Big)^2\Big]\bigg| \\
&\quad = O\big((B_S + rB_{\bar{K}})(B_S + rB_{\bar{K}} + rL_k B_k)|\alpha - \beta|\big).
\end{aligned}$$

By decomposing the interval $[0,1]$ into the disjoint union of intervals $((i-1)/N, i/N]$ for $i \in [N]$ and using this result, we can bound the first term of the right-hand side of inequality (55) as

$$\sup_{f\in\mathbb{B}(r,\bar{\mathcal{H}}),g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}),W\in\tilde{\mathcal{W}}} \big|\mathcal{R}(f,g,W)-\mathcal{R}_{\bar{\xi}}(f,g,W)\big| = O\bigg(\frac{(B_S+\bar{r}B_K)(B_S+\bar{r}B_K+\bar{r}L_K B_k)}{N}\bigg). \tag{57}$$

Eqn. (57) implies that we can transfer the results in Corollary 12 and Corollary 21 to $\mathcal{R}(\hat{f}_h,\hat{g}_h,\hat{W}_h)$ with an additional term shown in Eqn. (57). Thus, for the case where the MDP is induced by the distribution flow of the policy itself, we have that

$$\begin{aligned}&\mathcal{R}(\hat{f}_h,\hat{g}_h,\hat{W}_h)-\mathcal{R}(f_h^*,g_h^*,W_h^*)\\&\quad=O\bigg(\frac{(B_S+\bar{r}B_K)(B_S+\bar{r}B_K+\bar{r}L_K B_k)}{N}+\frac{(B_S+\bar{r}B_K)^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}\\&\qquad+\frac{(B_S+\bar{r}B_K)\bar{r}L_K B_k}{\sqrt{NL}}\log\frac{NL\mathcal{N}_\infty(1/\sqrt{NL},\tilde{\mathcal{W}})}{\delta}\bigg).\end{aligned} \tag{58}$$

For the case where the MDP is induced by a pre-specified distribution flow, we have that

$$\begin{aligned}&\mathcal{R}(\hat{f}_h',\hat{g}_h',\hat{W}_h')-\mathcal{R}(f_h^*,g_h^*,W_h^*)\\&\quad=O\bigg(\frac{(B_S+\bar{r}B_K)(B_S+\bar{r}B_K+\bar{r}L_K B_k)}{N}+\frac{(B_S+\bar{r}B_K)^4}{NL}\log\frac{N_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\tilde{r}}}N_{\tilde{\mathcal{W}}}}{\delta}\bigg).\end{aligned} \tag{59}$$

Step 3: Bound the estimation error of distribution flow and action-value function estimate.

For the estimation error of the distribution flow $\hat{\mu}_t^{\mathcal{I}}$, we have the following proposition

Proposition 22 *Given two GMFGs (P^*,r^*,W^*) and $(\hat{P},\hat{r},\hat{W})$, for a policy $\pi^{\mathcal{I}}\in\tilde{\Pi}$, we define the distribution flows induced by this policy as $\mu^{\mathcal{I}}=\Gamma_2(\pi^{\mathcal{I}},W^*)$ and $\hat{\mu}^{\mathcal{I}}=\hat{\Gamma}_2(\pi^{\mathcal{I}},\hat{W})$. Assume that the transition kernels P^* and $\hat{P}$ are equivalently defined by f^* and $\hat{f}\in\mathbb{B}(r,\bar{\mathcal{H}})$ from Eqn. (2). Under Assumption 8, we have that*

$$\|\hat{\mu}_h^\alpha-\mu_h^\alpha\|_1\le H(1+rL_K L_\varepsilon B_k)^H\sum_{m=1}^H\int_0^1 e_m^{\pi,\beta}\mathrm{d}\beta+\sum_{m=1}^H e_m^{\pi,\alpha},$$

where $e_h^{\pi,\alpha}$ is defined as

$$\begin{aligned}e_h^{\pi,\alpha}&=L_\varepsilon\sqrt{\mathbb{E}_{\rho_h^\alpha}\Big[\big(\hat{f}_h\big(\omega_h^\alpha(\hat{W}_h)-f_h^*\big(\omega_h^\alpha(W_h^*)\big)\big)^2\Big]},\\\omega_h^\alpha(W)&=\int_0^1\int_{\mathcal{S}}W(\alpha,\beta)k\big(\cdot,(s_{\tau,h}^i,a_{\tau,h}^i,s)\big)\mu_h^\beta(s)\mathrm{d}s\mathrm{d}\beta,\\\rho_h^\alpha&=\mu_h^\alpha\times\pi_h^\alpha\ \text{for}\ \alpha\in\mathcal{I}.\end{aligned}$$

Proof [Proof of Proposition 22] See Appendix Q.3.2. ■

From the definition of risk in Eqn. (8), we have that

$$\mathcal{R}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}(f_h^*, g_h^*, W_h^*)$$
$$= \frac{1}{L}\sum_{\tau=1}^{L}\int_0^1 \mathbb{E}_{\rho_{\tau,h}^\alpha}\Big[\Big(f_h^*\big(\omega_{\tau,h}^\alpha(W_h^*)\big) - \hat{f}_h\big(\omega_{\tau,h}^\alpha(\hat{W}_h)\big)\Big)^2 + \Big(g_h^*\big(\omega_{\tau,h}^\alpha(W_h^*)\big) - \hat{g}_h\big(\omega_{\tau,h}^\alpha(\hat{W}_h)\big)\Big)^2\Big]\mathrm{d}\alpha.$$

Since we implement the same policy $\pi_t^{\mathcal{I}}$ for L times in Step 1 of Algorithm 2, $\rho_{\tau,h}^\alpha$ for $\tau \in [L]$ are the same. Thus, we have

$$d(\hat{\mu}_t^{\mathcal{I}}, \mu_t^{\mathcal{I}}) = \sum_{h=1}^{H}\int_0^1 \|\hat{\mu}_{t,h}^\alpha - \mu_{t,h}^\alpha\|_1 \mathrm{d}\alpha \leq C\sum_{h=1}^{H}\sqrt{\mathcal{R}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}(f_h^*, g_h^*, W_h^*)},$$

where $C > 0$ is a constant, and the inequality results from Proposition 22 and Hölder inequality. The right-hande side of this inequality will play the role of ε_μ in the proof of Theorem 4, which is bounded in Eqn. (58).

Next, we bound the estimation error of the action-value function.

Proposition 23 *Assume that we have two GMFGs (P^*, r^*, W^*) and $(\hat{P}, \hat{r}, \hat{W})$. For a policy $\pi^{\mathcal{I}} \in \tilde{\Pi}$, a behavior policy $\pi^{\mathrm{b},\mathcal{I}} \in \tilde{\Pi}$, and a distribution flow $\mu^{\mathcal{I}} \in \tilde{\Delta}$, we define the distribution flows induced by the behavior policy on the GMFG (P^*, r^*, W^*) with underlying distribution flow $\mu^{\mathcal{I}}$ as $\mu^{\mathrm{b},\mathcal{I}} = \Gamma_3(\pi^{\mathrm{b},\mathcal{I}}, \mu^{\mathcal{I}}, W^*)$. Assume that the transition kernels P^* and $\hat{P}$ are equivalently defined by f^* and $\hat{f} \in \mathbb{B}(r, \bar{\mathcal{H}})$ from Eqn. (2), and reward functions r^* and $\hat{r}$ are equivalently defined by g^* and $\hat{g} \in \mathbb{B}(\tilde{r}, \tilde{\mathcal{H}})$ from Eqn. (2). Assume that $\sup_{s\in\mathcal{S},a\in\mathcal{A},\alpha\in\mathcal{I},h\in[H]} \pi_h^\alpha(a\,|\,s)/\pi_h^{\mathrm{b},\alpha}(a\,|\,s) \leq C$. Under Assumption 8, we have that*

$$\mathbb{E}_{\rho_h^{\mathrm{b},\alpha}}\Big[\Big|\hat{Q}_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},\hat{W}) - Q_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},W^*)\Big|\Big]$$
$$\leq C^H\sum_{m=h}^{H}\sqrt{\mathbb{E}_{\rho_m^{\mathrm{b},\alpha}}\Big[\Big(\hat{g}_m\big(\omega_h^\alpha(\hat{W}_m)\big) - g_m^*\big(\omega_h^\alpha(W_m^*)\big)\Big)^2\Big]}$$
$$+ L_\varepsilon H(1+\lambda\log|\mathcal{A}|)C^H\sum_{m=h}^{H}\sqrt{\mathbb{E}_{\rho_m^{\mathrm{b},\alpha}}\Big[\Big(\hat{f}_m\big(\omega_h^\alpha(\hat{W}_m)\big) - f_m^*\big(\omega_h^\alpha(W_m^*)\big)\Big)^2\Big]},$$

where $\rho_h^{\mathrm{b},\alpha}$ is defined as $\rho_h^{\mathrm{b},\alpha} = \mu_h^{\mathrm{b},\alpha}\cdot\pi_h^{\mathrm{b},\alpha}$, $\hat{Q}_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},\hat{W})$ and $Q_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},W^*)$ are the action-value functions of policy $\pi^{\mathcal{I}}$ on two GMFGs, and

$$\omega_h^\alpha(W) = \int_0^1\int_{\mathcal{S}} W(\alpha,\beta)k\big(\cdot,(s_{\tau,h}^i, a_{\tau,h}^i, s)\big)\mu_h^\beta(s)\mathrm{d}s\mathrm{d}\beta.$$

Proof [Proof of Proposition 23] See Appendix Q.3.3. ■

Next, we will make use of Proposition 23 to bound the estimation error of action-value function. Here, we adopt a different method to bound term (I) defined in Step 1 of the proof of Theorem 4. From inequality (78), we have

$$\text{(I)} \leq \eta_{t+1}\Big|\big\langle Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\bar{\mu}_t^{\mathcal{I}},W^*) - \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}), p - \pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle\Big|$$
$$+ \eta_{t+1}\Big|\big\langle \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}), \hat{\pi}_{t+1,h}^\alpha(\cdot\,|\,s_h) - \pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle\Big|$$

$$\leq 2\eta_{t+1}\big\|Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\bar{\mu}_t^{\mathcal{I}},W^*)-Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*)\big\|_\infty+2\eta_{t+1}H(1+\lambda\log|\mathcal{A}|)\beta_{t+1}$$
$$+\eta_{t+1}\Big|\big\langle Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*)-\hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}),p-\pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle\Big|. \quad (60)$$

For the third term in the right-hand side of inequality (60), if $p=\bar{\pi}_{t,h}^{*,\mathcal{I}}(\cdot\,|\,s_h)$, we have that

$$\Big|\big\langle Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*)-\hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}),\bar{\pi}_{t,h}^{*,\mathcal{I}}(\cdot\,|\,s_h)-\pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle\Big|$$
$$=\Bigg|\sum_{a\in\mathcal{A}}\big[Q_h^{\lambda,\alpha}(s_h,a_h,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*)-\hat{Q}_h^{\lambda,\alpha}(s_h,a_h,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})\big]$$
$$\cdot\,\pi_{t,h}^{\mathrm{b},\alpha}(a_h\,|\,s_h)\cdot\frac{\bar{\pi}_{t,h}^{*,\mathcal{I}}(a_h\,|\,s_h)-\pi_{t+1,h}^\alpha(a_h\,|\,s_h)}{\pi_{t,h}^{\mathrm{b},\alpha}(a_h\,|\,s_h)}\Bigg|$$
$$\leq(C_\pi+C_\pi')\sum_{a\in\mathcal{A}}\big|Q_h^{\lambda,\alpha}(s_h,a_h,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*)-\hat{Q}_h^{\lambda,\alpha}(s_h,a_h,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})\big|\cdot\pi_{t,h}^{\mathrm{b},\alpha}(a_h\,|\,s_h), \quad (61)$$

where the inequality results from Assumption 9. We note that we can let $p=\bar{\pi}_{t,h}^{*,\mathcal{I}}(\cdot\,|\,s_h)$ in our whole proof, because we will use such bound to upper bound the right-hand side of inequality (22), which we take $p=\bar{\pi}_{t,h}^{*,\mathcal{I}}(\cdot\,|\,s_h)$ to prove. Now we can define a new $\Lambda_{t+1,h}^\alpha$ with the terms in inequalities (60) and (61) replacing the original upper bound of term (I). In such case, the term ε_Q in inequality (27) can be replaced by the upper bound of the expectation of the third term in right-hand side in inequality (60).

$$\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\Bigg[\sum_{h=1}^H(C_\pi+C_\pi')\sum_{a\in\mathcal{A}}\big|Q_h^{\lambda,\alpha}(s_h,a_h,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*)-\hat{Q}_h^{\lambda,\alpha}(s_h,a_h,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})\big|\cdot\pi_{t,h}^{\mathrm{b},\alpha}(a_h\,|\,s_h)\Bigg]$$
$$\leq(C_\pi+C_\pi')C_\pi''\big(1+L_\varepsilon H(1+\lambda\log|\mathcal{A}|)\big)C_\pi^H H\sum_{h=1}^H\sqrt{\mathcal{R}(\hat{f}_h',\hat{g}_h',\hat{W}_h')-\mathcal{R}(f_h^*,g_h^*,W_h^*)},$$

where the inequality results from Propositions 27 and 23 and Assumption 10. The right-hand side of this inequality can be further bounded with Eqn. (59)

Step 4: Conclude the final result.

Replacing ε_μ and ε_Q with the derived new bounds and using the union bound, we have that

$$D\bigg(\frac{1}{T}\sum_{t=1}^T\pi_t^{\mathcal{I}},\pi^{*,\mathcal{I}}\bigg)+d\bigg(\frac{1}{T}\sum_{t=1}^T\hat{\bar{\mu}}_t^{\mathcal{I}},\mu^{*,\mathcal{I}}\bigg)$$
$$=O\bigg(\frac{\sqrt{\log T}}{T^{1/3}}\bigg)+O\bigg(\frac{(B_S+\bar{r}B_K)^{1/4}(\bar{r}L_KB_k)^{1/4}}{(NL)^{1/8}}\log^{1/4}\frac{TNL\mathcal{N}_\infty(1/\sqrt{N},\tilde{\mathcal{W}})}{\delta}$$
$$+\frac{B_S+\bar{r}B_K}{(NL)^{1/4}}\log^{1/4}\frac{TN_{\mathbb{B}_r}N_{\tilde{\mathbb{B}}_{\bar{r}}}N_{\tilde{\mathcal{W}}}}{\delta}+\frac{(B_S+\bar{r}B_K)^{1/4}(B_S+\bar{r}B_K+\bar{r}L_KB_k)^{1/4}}{N^{1/4}}\bigg).$$

Thus, we conclude the proof of Corollary 9. ∎

Appendix L. Proof of Corollary 10

Proof [Proof of Corollary 10] The proof of Corollary 10 is same as the proof of Corollary 9, except that we use the bound in Theorem 6 instead of Theorem 5. ■

Appendix M. Proof of Corollary 8

Proof [Proof of Corollary 8] We first define the inverse function of ψ^* as ϕ^*, i.e., $\phi^*(\psi^*(\alpha)) = \alpha$ for all $\alpha \in \mathcal{I}$. Similar to the proof of Theorem 7, we can decompose the risk difference as

$$\begin{aligned}&\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h^{\hat{\phi}_h \circ \psi^*}) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*)\\ &\quad = \text{Generalization Error of Risk} + \text{Estimation Error of Mean-embedding}\\ &\qquad + \text{Empirical Risk Difference}.\end{aligned}$$

For ease of notation, we only write the definition of each term for the transition kernel. The term for the reward functions can be easily derived.

$$\begin{aligned}&\text{Generalization Error of Risk}\\ &= \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\Big[\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f_h^*\big(\omega^i_{\tau,h}(W_h^*)\big)\Big)^2\Big]\\ &\quad - 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f_h^*\big(\omega^i_{\tau,h}(W_h^*)\big)\Big)^2\\ &\text{Estimation Error of Mean-embedding}\\ &= 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - \hat{f}_h\big(\hat{\bar{\omega}}^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h})\big)\Big)^2\\ &\quad + 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - f_h^*\big(\hat{\bar{\omega}}^i_{\tau,h}(W_h^{*,\phi^*})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f_h^*\big(\omega^i_{\tau,h}(W_h^*)\big)\Big)^2\\ &\text{Empirical Risk Difference}\\ &= 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\hat{\bar{\omega}}^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f_h^*\big(\hat{\bar{\omega}}^i_{\tau,h}(W_h^{*,\phi^*})\big)\Big)^2.\end{aligned}$$

From the estimation procedure of Algorithm 10, we have that

$$\text{Empirical Risk Difference} \le 0.$$

For Estimation Error of Mean-embedding, we can use the bound in inequality (50) in the proof of Theorem 7 to bound it. In fact, since ψ^* is the inverse function of ϕ^*, the expression of the Estimation Error of Mean-embedding here is same as the term in inequality (50). For generalization error of risk, we can use inequality (54) in the proof of Theorem 7 to bound it. Thus, we conclude the proof of Corollary 8. ■

Appendix N. Proof of Corollary 11

Proof [Proof of Corollary 11] We note that Line 6 of Algorithm 2 involves the estimation of MDP when the underlying distribution flow is given. However, different from the setting in Section 6.1, here we can only specify the distribution flow through $\{\mu^{(\xi_i-1/N,\xi_i]}\}_{i=1}^N$. We concatenate these distribution flows to form $\tilde{\mu}^{\mathcal{I}}$ that is defined as $\tilde{\mu}^\alpha = \mu^{\xi_i+\alpha-i/N}$ if $\alpha \in ((i-1)/N, i/N]$. That is, we assume that $\xi_i = i/N$. Then we define the mean-embedding induced by $\tilde{\mu}^{\mathcal{I}}$ as

$$\tilde{\omega}^i_{\tau,h}(W) = \int_0^1 \int_{\mathcal{S}} W(i/N, \beta) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big) \tilde{\mu}^\beta_{\tau,h}(s) \mathrm{d}s \mathrm{d}\beta.$$

Then we estimate the transition kernels, reward functions, and graphons as

$$(\hat{f}_h, \hat{g}_h, \hat{W}_h, \hat{\phi}_h) = \mathop{\mathrm{argmin}}_{\substack{f\in\mathbb{B}(r,\tilde{\mathcal{H}}),\\ g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}),\\ W\in\tilde{\mathcal{W}},\\ \phi\in\mathcal{C}^N_{[0,1]}}} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s^i_{\tau,h+1} - f\big(\tilde{\omega}^i_{\tau,h}(W^\phi)\big)\Big)^2 + \Big(r^i_{\tau,h} - g\big(\tilde{\omega}^i_{\tau,h}(W^\phi)\big)\Big)^2. \tag{62}$$

Here we adopt the similar steps as the proof of Corollary 9. We note that the only different procedure is the first step. Next, we will derive the performance guarantee of Algorithm (62).

In such setting, we implement $\{\pi^{(\xi_i-1/N,\xi_i]}\}_{i=1}^N$ for L times on the MDP induced by $\{\mu^{(\xi_i-1/N,\xi_i]}\}_{i=1}^N$ to collect the dataset $\mathcal{D}_\tau = \{(s^{[N]}_{\tau,h}, a^{[N]}_{\tau,h}, r^{[N]}_{\tau,h}, s^{[N]}_{\tau,h+1})\}_{h=1}^H$ for $\tau \in [L]$. We define $\mu^{+,\mathcal{I}} = \Gamma_3(\pi^{\mathcal{I}}, \mu^{\mathcal{I}}, W^*)$ as the distribution flow of implementing $\pi^{\mathcal{I}}$ on the MDP induced by $\mu^{\mathcal{I}}$. We highlight that we will not use quantity in the estimation procedure, but use it only in the analysis. The joint distribution of $(s^i_{\tau,h}, a^i_{\tau,h}, r^i_{\tau,h}, s^i_{\tau,h+1})_{i=1}^N$ is $\prod_{i=1}^N \rho^{+,i}_{\tau,h}$, where $\rho^{+,i}_{\tau,h} = \mu^{+,i}_{\tau,h} \times \pi^i_{\tau,h} \times \delta_{r^*_h} \times P^*_h$. Same as the proof of Corollary 8, we define two bijections $\psi^*, \phi^* \in \mathcal{C}^N_{[0,1]}$ as $\psi^*(\xi_i) = i/N$ for all $i \in [N]$, and $\phi^* \circ \psi^*(\alpha) = \phi^*(\psi^*(\alpha)) = \alpha$ for all $\alpha \in \mathcal{I}$.

With a little abuse of notation, we define the risk of (f, g, W) given $\bar{\xi}$ as

$$\mathcal{R}_{\bar{\xi}}(f, g, W) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\rho^{+,i}_h}\Big[\Big(s^i_{h+1} - f\big(\omega^i_h(W)\big)\Big)^2 + \Big(r^i_{h+1} - g\big(\omega^i_h(W)\big)\Big)^2\Big].$$

Corollary 24 *Under Assumptions 5, 6, 7, and 1, if $\{\xi_i\}_{i=1}^N = \{i/N\}_{i=1}^N$, then the risk of estimate derived in Algorithm* (62) *can be bounded as*

$$\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h^{\hat{\phi}_h\circ\psi^*}) - \mathcal{R}_{\bar{\xi}}(f^*_h, g^*_h, W^*_h) \le O\bigg(\frac{(B_S + rB_{\bar{K}})^4}{L} \log \frac{N\tilde{N}_{\mathbb{B}_r}\tilde{N}_\infty}{\delta}\bigg).$$

with probability at least $1-\delta$.

Proof [Proof of Corollary 24] See Appendix Q.3.4. ■

Then we only need to exactly follow the steps 2, 3, and 4 in the proof of Corollary 9 to prove the desired results. Thus, we conclude the proof of Corollary 11. ■

Appendix O. Proof of Proposition 3

The existence follows from the Banach fixed point theorem. To prove the uniqueness, we adopt proof by contradiction. In this case, we admit the existence of multiple different NEs. Then the expectations in distances $D(\cdot,\cdot)$ and $d(\cdot,\cdot)$ defined in Section 4.2 can be taken with respect to any NE.

Proof [Proof of Proposition 3] We first prove the existence of NE under Assumption 2. From the mean-field side, for two mean fields $\mu^{\mathcal{I}}$ and $\tilde{\mu}^{\mathcal{I}}$, we have that

$$\begin{aligned} & d\Big(\Gamma_2\big(\Gamma_1^{\lambda}(\mu^{\mathcal{I}}, W^*), W^*\big), \Gamma_2\big(\Gamma_1^{\lambda}(\tilde{\mu}^{\mathcal{I}}, W^*), W^*\big)\Big) \\ & \quad \le d_2 \cdot D\big(\Gamma_1^{\lambda}(\mu^{\mathcal{I}}, W^*), \Gamma_1^{\lambda}(\tilde{\mu}^{\mathcal{I}}, W^*)\big) \\ & \quad \le d_1 d_2 d(\mu^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}}). \end{aligned}$$

Banach fixed point theorem shows that there exists a fixed point for $\Gamma_2 \circ \Gamma_1^{\lambda}$, which we denote as $\mu^{*,\mathcal{I}}$. We then define $\pi^{*,\mathcal{I}} = \Gamma_1^{\lambda}(\mu^{*,\mathcal{I}}, W^*)$. Definition 1 shows that $(\pi^{*,\mathcal{I}}, \mu^{*,\mathcal{I}})$ is a NE.

Then we prove the uniqueness of NE. Assume that there are two NEs $(\pi^{*,\mathcal{I}}, \mu^{*,\mathcal{I}})$ and $(\tilde{\pi}^{*,\mathcal{I}}, \tilde{\mu}^{*,\mathcal{I}})$. From the Definition 1 of the NE, we have that

$$\pi^{*,\mathcal{I}} = \Gamma_1^{\lambda}(\mu^{*,\mathcal{I}}, W^*), \quad \mu^{*,\mathcal{I}} = \Gamma_2(\pi^{*,\mathcal{I}}, W^*), \quad \tilde{\pi}^{*,\mathcal{I}} = \Gamma_1^{\lambda}(\tilde{\mu}^{*,\mathcal{I}}, W^*), \quad \tilde{\mu}^{*,\mathcal{I}} = \Gamma_2(\tilde{\pi}^{*,\mathcal{I}}, W^*).$$

Then Assumption 2 implies that

$$d(\mu^{*,\mathcal{I}}, \tilde{\mu}^{*,\mathcal{I}}) \le d_1 d_2 d(\mu^{*,\mathcal{I}}, \tilde{\mu}^{*,\mathcal{I}}).$$

Thus, we have $d(\mu^{*,\mathcal{I}}, \tilde{\mu}^{*,\mathcal{I}}) = 0$, which implies that they are different only on a set of zero-measure agents with respect to the Lebesgue measure on $[0,1]$. Thus, we conclude the proof of Proposition 3. ∎

Appendix P. Lipschitzness of NE

Proposition 25 *Under Assumptions 1 and 5, for any NE of the λ-regularized GMFG $(\pi^{\lambda,\mathcal{I}}, \mu^{\lambda,\mathcal{I}})$ with $\lambda > 0$, we have that*

$$\big\|\pi_h^{\lambda,\alpha}(\cdot \,|\, s) - \pi_h^{\lambda,\beta}(\cdot \,|\, s)\big\|_1 \le \frac{2HL_{\mathcal{W}}\Big[L_r + H\big(1 + \lambda \log|\mathcal{A}|\big)L_P\Big]}{\lambda}|\alpha - \beta| \text{ for all } h \in [H], s \in \mathcal{S}.$$

Proof [Proof of Proposition 25] For any distribution flow $\mu^{\mathcal{I}}$, we denote the optimal value function in the λ-regularized MDP induced by $\mu^{\mathcal{I}}$ as $V^{*,\mathcal{I}} = (V_h^{*,\mathcal{I}})_{h=1}^H$. Then we prove the proposition in two steps:

- Given any distribution flow $\mu^{\mathcal{I}}$, the optimal value function $V^{*,\mathcal{I}}$ is Lipschitz in the positions of agents, i.e., $|V_h^{*,\alpha}(s) - V_h^{*,\beta}(s)| \le H[L_r + H(1 + \lambda \log|\mathcal{A}|)L_P]L_{\mathcal{W}}|\alpha - \beta|$ for all $s \in \mathcal{S}$ and $h \in [H]$.
- Any policy $\pi^{\mathcal{I}}$ that achieves the optimal value function $V^{*,\mathcal{I}}$ is Lipschitz in the positions of agents.

These two steps concludes the proof of Proposition 25 by noting that for any λ-NE $(\pi^{\lambda,\mathcal{I}}, \mu^{\lambda,\mathcal{I}})$, the policy $\pi^{\lambda,\mathcal{I}}$ achieves the maximal accumulative rewards in the MDP induced by $\mu^{\lambda,\mathcal{I}}$ according to Definition 1.

Step 1: Show the optimal value function $V^{*,\mathcal{I}}$ is Lipschitz in the positions of agents.

For any distribution flow $\mu^{\mathcal{I}} \in \Delta(\mathcal{S})^{\mathcal{I}\times H}$, we define an operator acting on $\mathcal{S} \to \mathbb{R}$ as

$$T_h^{\mu^{\mathcal{I}},\alpha} u(s) = \sup_{p\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} p(a) r_h(s,a,z_h^\alpha) - \lambda R(p) + \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} p(a) P_h(s' \mid s,a,z_h^\alpha) u(s') \mathrm{d}s' \text{ for } h \in [H-1],$$
$$T_H^{\mu^{\mathcal{I}},\alpha} u(s) = \sup_{p\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} p(a) r_H(s,a,z_H^\alpha) - \lambda R(p),$$

where $R(\cdot)$ is the negative entropy function. Since $V^{*,\mathcal{I}}$ is the optimal value function of the MDP induced by $\mu^{\mathcal{I}}$, we have that

$$T_h^{\mu^{\mathcal{I}},\alpha} V_{h+1}^{*,\alpha}(s) = V_h^{*,\alpha}(s) \text{ and } V_{H+1}^{*,\alpha}(s) = 0 \text{ for all } s \in \mathcal{S}, h \in [H], \alpha \in \mathcal{I}.$$

For any $h \in [H-1]$, we have that

$$\begin{aligned}
&\big|V_h^{*,\alpha}(s) - V_h^{*,\beta}(s)\big| \\
&\quad \le \sup_{p\in\Delta(\mathcal{A})} \bigg| \sum_{a\in\mathcal{A}} p(a)\big(r_h(s,a,z_h^\alpha) - r_h(s,a,z_h^\beta)\big) \\
&\qquad + \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} p(a)\big(P_h(s' \mid s,a,z_h^\alpha) V_{h+1}^{*,\alpha}(s') - P_h(s' \mid s,a,z_h^\beta) V_{h+1}^{*,\beta}(s')\big) \mathrm{d}s' \bigg| \\
&\quad \le L_r \|z_h^\alpha - z_h^\beta\|_1 + H(1+\lambda\log|\mathcal{A}|) L_P \|z_h^\alpha - z_h^\beta\|_1 + \sup_{s\in\mathcal{S}} \big|V_{h+1}^{*,\alpha}(s) - V_{h+1}^{*,\beta}(s)\big|,
\end{aligned}$$

where the first inequality results from Assumption 1. Note that $\|z_h^\alpha - z_h^\beta\|_1 \le \| \int_0^1 \big(W_h(\alpha,\gamma) - W_h(\beta,\gamma)\big)\mu_h^\gamma \mathrm{d}\gamma\|_1 \le L_{\mathcal{W}}|\alpha-\beta|$, we have

$$\sup_{s\in\mathcal{S}} \big|V_h^{*,\alpha}(s) - V_h^{*,\beta}(s)\big| \le \Big[L_r + H\big(1+\lambda\log|\mathcal{A}|\big)L_P\Big] L_{\mathcal{W}}|\alpha-\beta| + \sup_{s\in\mathcal{S}} \big|V_{h+1}^{*,\alpha}(s) - V_{h+1}^{*,\beta}(s)\big|.$$

Summing this inequality for $t = h, \cdots, H$ and noting that $V_{H+1}^{*,\alpha}(s) = 0$, we have

$$\sup_{s\in\mathcal{S}} \big|V_h^{*,\alpha}(s) - V_h^{*,\beta}(s)\big| \le H\Big[L_r + H\big(1+\lambda\log|\mathcal{A}|\big)L_P\Big] L_{\mathcal{W}}|\alpha-\beta|.$$

Step 2: Any policy that achieves the optimal value function $V^{*,\mathcal{I}}$ is Lipschitz in the positions of agents

Assume that policy $\pi^{\mathcal{I}}$ achieves the optimal value function $V^{*,\mathcal{I}}$. For any $\alpha, \beta \in \mathcal{I}$, $s \in \mathcal{S}$, and $h \in [H]$, we have that

$$\pi_h^\alpha(\cdot \mid s) = \operatorname*{argmax}_{p\in\Delta(\mathcal{A})} \sum_{a\in\mathcal{A}} p(a) r_h(s,a,z_h^\alpha) - \lambda R(p) + \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} p(a) P_h(s' \mid s,a,z_h^\alpha) V_{h+1}^{*,\alpha}(s') \mathrm{d}s'$$

$$\pi_h^\beta(\cdot \mid s) = \operatorname*{argmax}_{p \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) r_h(s, a, z_h^\beta) - \lambda R(p) + \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} p(a) P_h(s' \mid s, a, z_h^\beta) V_{h+1}^{*,\beta}(s') \mathrm{d}s'$$

Define $y^\alpha(s, a) = r_h(s, a, z_h^\alpha) + \int_{\mathcal{S}} p(a) P_h(s' \mid s, a, z_h^\alpha) V_{h+1}^{*,\alpha}(s') \mathrm{d}s'$ for all $\alpha \in \mathcal{I}$. Lemma 34 shows that

$$\big\|\pi_h^\alpha(\cdot \mid s) - \pi_h^\beta(\cdot \mid s)\big\|_1 \le \frac{1}{\lambda} \|y^\alpha(s, \cdot) - y^\beta(s, \cdot)\|_\infty.$$

Term $\|y^\alpha(s, \cdot) - y^\beta(s, \cdot)\|_\infty$ can be bounded as

$$\begin{aligned}
&\|y^\alpha(s, \cdot) - y^\beta(s, \cdot)\|_\infty \\
&\quad \le \Big[L_r + H\big(1 + \lambda \log|\mathcal{A}|\big) L_P\Big] L_{\mathcal{W}} |\alpha - \beta| + H\Big[L_r + H\big(1 + \lambda \log|\mathcal{A}|\big) L_P\Big] L_{\mathcal{W}} |\alpha - \beta| \\
&\quad \le 2H\Big[L_r + H\big(1 + \lambda \log|\mathcal{A}|\big) L_P\Big] L_{\mathcal{W}} |\alpha - \beta|,
\end{aligned}$$

where the first inequality results from the triangle inequality, and the second inequality results from Step 1. Thus, we conclude that

$$\big\|\pi_h^\alpha(\cdot \mid s) - \pi_h^\beta(\cdot \mid s)\big\|_1 \le \frac{2H\Big[L_r + H\big(1 + \lambda \log|\mathcal{A}|\big) L_P\Big] L_{\mathcal{W}}}{\lambda} |\alpha - \beta|,$$

which proves the claim of Proposition 25. ■

Appendix Q. Supporting Propositions and Lemmas

Q.1 Propositions and Lemmas for Estimation

Q.1.1 Proof of Proposition 19

Proof [Proof of Proposition 19] For any $h \in [H-1]$, we have that

$$\begin{aligned}
&\|\mu_{h+1}^\alpha - \mu_{h+1}^\beta\|_1 \\
&\quad = \int_{\mathcal{S}} \bigg| \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} P_h(s' \mid s, a, z_h^\alpha) \mu_h^\alpha(s) \pi_h^\alpha(a \mid s) \mathrm{d}s - \sum_{a \in \mathcal{A}} \int_{\mathcal{S}} P_h(s' \mid s, a, z_h^\beta) \mu_h^\beta(s) \pi_h^\beta(a \mid s) \mathrm{d}s \bigg| \mathrm{d}s' \\
&\quad \le L_P \|z_h^\alpha - z_h^\beta\|_1 + \|\mu_h^\alpha - \mu_h^\beta\|_1 + \sup_{s \in \mathcal{S}} \big\|\pi_h^\alpha(\cdot \mid s) - \pi_h^\beta(\cdot \mid s)\big\|_1, \qquad (63)
\end{aligned}$$

where the first inequality results from the triangle inequality, and the second inequality results from Assumptions 5 and 1. We further bound the first term in the right-hand side of inequality (63) as

$$\|z_h^\alpha - z_h^\beta\|_1 = \bigg\| \int_0^1 W_h(\alpha, \gamma) \mu_h^\gamma \mathrm{d}\gamma - \int_0^1 W_h(\beta, \gamma) \mu_h^\gamma \mathrm{d}\gamma \bigg\|_1 \le L_{\mathcal{W}} |\alpha - \beta|,$$

where the inequality results from Assumption 5. Substituting this inequality to the right-hand side of inequality (63), we derive that

$$\|\mu_{h+1}^\alpha - \mu_{h+1}^\beta\|_1 \le \|\mu_h^\alpha - \mu_h^\beta\|_1 + L_P L_{\mathcal{W}} |\alpha - \beta| + \sup_{s \in \mathcal{S}} \big\|\pi_h^\alpha(\cdot \mid s) - \pi_h^\beta(\cdot \mid s)\big\|_1.$$

Summing these inequalities for $h = 1, \cdots, t$, we have that

$$\|\mu_t^\alpha - \mu_t^\beta\|_1 \le (t-1)L_P L_{\mathcal{W}}|\alpha - \beta| + \sum_{h=1}^{t-1} \sup_{s\in\mathcal{S}} \big\|\pi_h^\alpha(\cdot\,|\,s) - \pi_h^\beta(\cdot\,|\,s)\big\|_1,$$

which results from that $\mu_1^\alpha = \mu_1^\beta$. Thus, we concludes the proof of Proposition 19. ■

Q.1.2 Proof of Proposition 20

Proof [Proof of Proposition 20] Our proof of Proposition 20 follows the pipeline of the proof of Györfi et al. (2002, Theorem 11.4). However, the random variables in our problem are not identically distributed, which requires additional techniques to control the tail probabilities. Our proof involves three steps:

- Symmetrization by a ghost sample.
- Additional randomization by random signs.
- Bounding the covering number

Step 1: Symmetrization by a ghost sample.

We construct the random variables $\tilde{D}_h = \{\tilde{e}_{\tau,h}^i\}_{\tau,i=1}^{L,N}$ that are independent of and identically distributed as $D_h = \{e_{\tau,h}^i\}_{\tau,i=1}^{L,N}$. It means that $\tilde{e}_{\tau,h}^i \overset{D}{=} e_{\tau,h}^i$ for all $\tau \in [L], i \in [N]$, and they are independent. For ease of notation, we write $\sum_{\tau=1}^L \sum_{i=1}^N$ as $\sum_{\tau,i}$ Choose a function f_W that depends on D_h such that

$$\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(e_{\tau,h}^i)\big] - \frac{1}{NL}\sum_{\tau,i} f_W(e_{\tau,h}^i) \ge \varepsilon\bigg(\alpha + \beta + \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(e_{\tau,h}^i)\big]\bigg)$$

holds. If such function does not exist, then f_W is an arbitrary function in $\mathcal{F}_{\tilde{\mathcal{W}}}$. Then we have that

$$\begin{aligned}
&\mathbb{E}_{\rho_{\tau,h}^i}\bigg[\Big(f_W(\tilde{e}_{\tau,h}^i) - \mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(\tilde{e}_{\tau,h}^i)\,\big|\,D_h\big]\Big)^2\,\bigg|\,D_h\bigg]\\
&\quad\le \mathbb{E}_{\rho_{\tau,h}^i}\Big[\big(f_W(\tilde{e}_{\tau,h}^i)\big)^2\,\Big|\,D_h\Big]\\
&\quad\le 4(B_S + rB_{\bar{K}})^2\mathbb{E}_{\rho_{\tau,h}^i}\bigg[\Big(f\big(\tilde{\omega}_{\tau,h}^i(W)\big) - f_h^*\big(\tilde{\omega}_{\tau,h}^i(W_h^*)\big)\Big)^2\,\bigg|\,D_h\bigg]\\
&\quad= 4(B_S + rB_{\bar{K}})^2\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(\tilde{e}_{\tau,h}^i)\,\big|\,D_h\big],
\end{aligned} \tag{64}$$

where the second inequality results from Lemma 33, and the last equality results from that $\mathbb{E}_{\rho_{\tau,h}^i}[\tilde{s}_{\tau,h+1}^i\,|\,D_h, \tilde{\omega}_{\tau,h}^i(W_h^*)] = f_h^*(\tilde{\omega}_{\tau,h}^i(W_h^*))$. Then the tail probability for the ghost sample $\tilde{D}_h$ is bouneded as

$$\mathbb{P}\bigg(\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(\tilde{e}_{\tau,h}^i)\,\big|\,D_h\big] - \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}_{\tau,h}^i) \ge \frac{\varepsilon}{2}\bigg(\alpha + \beta + \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho_{\tau,h}^i}\big[f_W(\tilde{e}_{\tau,h}^i)\,\big|\,D_h\big]\bigg)\bigg)$$

$$\leq \frac{\mathbb{E}\Big[\Big(\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\,\big|\,D_h\big] - \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}^i_{\tau,h})\Big)^2\Big]}{\Big(\frac{\varepsilon(\alpha+\beta)}{2} + \frac{\varepsilon}{2NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\,\big|\,D_h\big]\Big)^2}$$
$$\leq \frac{\frac{4(B_S+rB_{\bar{K}})^2}{NL}\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\,\big|\,D_h\big]}{\Big(\frac{\varepsilon(\alpha+\beta)}{2} + \frac{\varepsilon}{2NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\,\big|\,D_h\big]\Big)^2}$$
$$\leq \frac{4(B_S+rB_{\bar{K}})^2}{(\alpha+\beta)NL\varepsilon^2},$$

where the first inequality results from Chebyshev inequality, the second inequality results from inequality (64), and the last inequality results from $x/(a+x)^2 \leq 1/(4a)$ for any $x, a > 0$. When $NL \geq 32(B_S + rB_{\bar{K}})^2/((\alpha+\beta)\varepsilon^2)$, we have that

$$\mathbb{P}\bigg(\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\,\big|\,D_h\big] - \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}^i_{\tau,h}) \geq \frac{\varepsilon}{2}\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\,\big|\,D_h\big]\bigg)\bigg) \leq \frac{1}{8}. \tag{65}$$

Thus, we have that

$$\begin{aligned}
&\mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big] - \frac{1}{NL}\sum_{\tau,i} f_W(e^i_{\tau,h}) \geq \varepsilon\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]\bigg)\bigg)\\
&\leq \frac{8}{7}\mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\big] - f_W(\tilde{e}^i_{\tau,h}) < \frac{\varepsilon}{2}\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(\tilde{e}^i_{\tau,h})\big]\bigg),\\
&\qquad \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big] - f_W(e^i_{\tau,h}) \geq \varepsilon\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]\bigg)\bigg)\\
&\leq \frac{8}{7}\mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}^i_{\tau,h}) - f_W(e^i_{\tau,h}) \geq \frac{\varepsilon}{2}\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]\bigg)\bigg),
\end{aligned} \tag{66}$$

where the first inequality results from inequality (65), the detailed proof of this step is in Györfi et al. (2002, Theorem 11.4). To derive the fast rate result, we want to replace the expectation of the function in the right-hand side of inequality (66) by the expectation of the square of the function. Thus, We handle the right-hand side of inequality (66) as

$$\begin{aligned}
&\mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}^i_{\tau,h}) - \frac{1}{NL}\sum_{\tau,i} f_W(e^i_{\tau,h}) \geq \frac{\varepsilon}{2}\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]\bigg)\bigg)\\
&\leq \mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}^i_{\tau,h}) - \frac{1}{NL}\sum_{\tau,i} f_W(e^i_{\tau,h}) \geq \frac{\varepsilon}{2}\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]\bigg),\\
&\qquad \frac{1}{NL}\sum_{\tau,i} f^2_W(e^i_{\tau,h}) - \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f^2_W(e^i_{\tau,h})\big] \leq \varepsilon\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f^2_W(e^i_{\tau,h})\big]\bigg),
\end{aligned}$$

$$\frac{1}{NL}\sum_{\tau,i} f_W^2(\tilde{e}^i_{\tau,h}) - \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(\tilde{e}^i_{\tau,h})\big] \le \varepsilon\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(\tilde{e}^i_{\tau,h})\big]\bigg)\bigg)$$
$$+ 2\mathbb{P}\bigg(\exists f_W\in\mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i} f_W^2(\tilde{e}^i_{\tau,h}) - \mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(\tilde{e}^i_{\tau,h})\big] \le \varepsilon\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(\tilde{e}^i_{\tau,h})\big]\bigg)\bigg)$$
$$= (\mathrm{VII}) + (\mathrm{VIII}), \tag{67}$$

where the inequality follows from the union bound. For the term (VIII), Proposition 26 shows that

$$\mathbb{P}\bigg(\exists f_W\in\mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i} f_W^2(\tilde{e}^i_{\tau,h}) - \frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(\tilde{e}^i_{\tau,h})\big] \le \varepsilon\bigg(\alpha+\beta$$
$$+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(\tilde{e}^i_{\tau,h})\big]\bigg)\bigg)$$
$$\le 2\mathbb{E}\bigg[\mathcal{N}_1\bigg(\frac{\varepsilon(\alpha+\beta)}{5}, \{f_W^2 \mid f_W\in\mathcal{F}_{\tilde{\mathcal{W}}}\}, \{\tilde{e}^i_{\tau,h}\}_{\tau,i=1}^{L,N}\bigg)\bigg]\exp\bigg(-\frac{3\varepsilon^2(\alpha+\beta)NL}{40(B_S+rB_{\bar{K}})^4}\bigg)$$
$$\le 2\mathbb{E}\bigg[\mathcal{N}_1\bigg(\frac{\varepsilon(\alpha+\beta)}{10(B_S+rB_{\bar{K}})^2}, \mathcal{F}_{\tilde{\mathcal{W}}}, \{\tilde{e}^i_{\tau,h}\}_{\tau,i=1}^{L,N}\bigg)\bigg]\exp\bigg(-\frac{3\varepsilon^2(\alpha+\beta)NL}{40(B_S+rB_{\bar{K}})^4}\bigg). \tag{68}$$

For term (VII), the last two events in (VII) are equivalent to that

$$(1+\varepsilon)\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(e^i_{\tau,h})\big] \ge (1-\varepsilon)\frac{1}{NL}\sum_{\tau,i} f_W^2(e^i_{\tau,h}) - \varepsilon(\alpha+\beta)$$
$$(1+\varepsilon)\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W^2(\tilde{e}^i_{\tau,h})\big] \ge (1-\varepsilon)\frac{1}{NL}\sum_{\tau,i} f_W^2(\tilde{e}^i_{\tau,h}) - \varepsilon(\alpha+\beta). \tag{69}$$

Then term (VII) can be bounded as

$$(\mathrm{VII})$$
$$\le \mathbb{P}\bigg(\exists f_W\in\mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}^i_{\tau,h}) - \frac{1}{NL}\sum_{\tau,i} f_W(e^i_{\tau,h}) \ge \frac{\varepsilon(\alpha+\beta)}{2}$$
$$-\frac{\varepsilon^2(\alpha+\beta)}{4(B_S+rB_{\bar{K}})^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{8(B_S+rB_{\bar{K}})^2(1+\varepsilon)}\frac{1}{NL}\sum_{\tau,i} f_W^2(e^i_{\tau,h}) + f_W^2(\tilde{e}^i_{\tau,h})\bigg), \tag{70}$$

where the inequality results from inequality (69) and the fact that $\mathbb{E}_{\rho^i_{\tau,h}}[f_W^2(e^i_{\tau,h})] \le 4(B_S + rB_{\bar{K}})^2\mathbb{E}_{\rho^i_{\tau,h}}[f_W(e^i_{\tau,h})]$.

Proposition 26 *Let $B\ge 1$, $\mathcal{G}$ be a set of functions $g:\mathcal{X}\to[0,B]$. Let $Z_1,\cdots,Z_n$ be independent $\mathcal{X}-$valued random variables that are distributed as $\rho_1,\cdots,\rho_n$, respectively. Assume $\alpha>0$, $0<\varepsilon\le 1$, $n\ge 1$. Then we have that*

$$\mathbb{P}\bigg(\sup_{g\in\mathcal{G}}\frac{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\rho_i}\big[g(Z)\big]}{\alpha+\frac{1}{n}\sum_{i=1}^n g(Z_i)+\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\rho_i}\big[g(Z)\big]} > \varepsilon\bigg) \le 2\mathbb{E}\bigg[\mathcal{N}_1\bigg(\frac{\alpha\varepsilon}{5},\mathcal{G},Z_1^n\bigg)\bigg]\exp\bigg(-\frac{3\alpha n\varepsilon^2}{40B}\bigg)$$

for $n\ge 16B/(\varepsilon^2\alpha)$.

Proof [Proof of Proposition 26] See Appendix Q.1.3. ■

Step 2: Additional randomization by random signs.

Let $\{U_{\tau,h}^i\}_{\tau,i=1}^{L,N}$ be independent and uniformly distributed over $\{+1,-1\}$ that are also independent of D_h and $\tilde{D}_h$. Then we have that

$$\begin{aligned}
&\mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \frac{1}{NL}\sum_{\tau,i} f_W(\tilde{e}_{\tau,h}^i) - \frac{1}{NL}\sum_{\tau,i} f_W(e_{\tau,h}^i) \geq \frac{\varepsilon(\alpha+\beta)}{2} \\
&\qquad - \frac{\varepsilon^2(\alpha+\beta)}{4(B_S + rB_{\bar{K}})^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{8(B_S+rB_{\bar{K}})^2(1+\varepsilon)}\frac{1}{NL}\sum_{\tau,i} f_W^2(e_{\tau,h}^i) + f_W^2(\tilde{e}_{\tau,h}^i)\bigg) \\
&\leq 2\mathbb{E}\bigg[\mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \bigg|\frac{1}{NL}\sum_{\tau,i} U_{\tau,h}^i f_W(e_{\tau,h}^i)\bigg| \geq \frac{\varepsilon(\alpha+\beta)}{4} \\
&\qquad - \frac{\varepsilon^2(\alpha+\beta)}{8(B_S + rB_{\bar{K}})^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{8(B_S+rB_{\bar{K}})^2(1+\varepsilon)}\frac{1}{NL}\sum_{\tau,i} f_W^2(e_{\tau,h}^i) \,\bigg|\, \{e_{\tau,h}^i\}_{\tau,i=1}^{L,N}\bigg)\bigg]
\end{aligned} \tag{71}$$

where the inequality results from the union bound. Let $\delta > 0$, $\mathcal{F}_\delta$ be a L_1 $\delta-$cover of $\mathcal{F}_{\tilde{\mathcal{W}}}$ on $\{e_{\tau,h}^i\}_{\tau,i=1}^{L,N}$. Then for any $f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}$, there exists $\bar{f}_W \in \mathcal{F}_\delta$ such that

$$\frac{1}{NL}\sum_{\tau,i}\left|f_W(e_{\tau,h}^i) - \bar{f}_W(e_{\tau,h}^i)\right| \leq \delta.$$

This inequality implies that

$$\begin{aligned}
&\bigg|\frac{1}{NL}\sum_{\tau,i} U_{\tau,h}^i f_W(e_{\tau,h}^i)\bigg| - \bigg|\frac{1}{NL}\sum_{\tau,i} U_{\tau,h}^i \bar{f}_W(e_{\tau,h}^i)\bigg| \leq \delta \\
&\qquad \frac{1}{NL}\sum_{\tau,i} f_W^2(e_{\tau,h}^i) - \frac{1}{NL}\sum_{\tau,i} \bar{f}_W^2(e_{\tau,h}^i) \geq -2(B_S + rB_{\bar{K}})^2\delta,
\end{aligned}$$

where these inequalities results from the triangle inequality. In the following, we take $\delta = \varepsilon\beta/5$. Thus, we can bound the right-hand side of inequality (71) as

$$\begin{aligned}
&\mathbb{P}\bigg(\exists f_W \in \mathcal{F}_{\tilde{\mathcal{W}}}, \bigg|\frac{1}{NL}\sum_{\tau,i} U_{\tau,h}^i f_W(e_{\tau,h}^i)\bigg| \geq \frac{\varepsilon(\alpha+\beta)}{4} \\
&\qquad - \frac{\varepsilon^2(\alpha+\beta)}{8(B_S + rB_{\bar{K}})^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{8(B_S+rB_{\bar{K}})^2(1+\varepsilon)}\frac{1}{NL}\sum_{\tau,i} f_W^2(e_{\tau,h}^i) \,\bigg|\, \{e_{\tau,h}^i\}_{\tau,i=1}^{L,N}\bigg) \\
&\leq \mathcal{N}_1\bigg(\frac{\varepsilon\beta}{5}, \mathcal{F}_{\tilde{\mathcal{W}}}, \{e_{\tau,h}^i\}_{\tau,i=1}^{L,N}\bigg) \max_{f_W \in \mathcal{F}_{\frac{\varepsilon\beta}{5}}} \mathbb{P}\bigg(\bigg|\frac{1}{NL}\sum_{\tau,i} U_{\tau,h}^i f_W(e_{\tau,h}^i)\bigg| \geq \frac{\varepsilon\alpha}{4} \\
&\qquad - \frac{\varepsilon^2\alpha}{8(B_S + rB_{\bar{K}})^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{8(B_S+rB_{\bar{K}})^2(1+\varepsilon)}\frac{1}{NL}\sum_{\tau,i} f_W^2(e_{\tau,h}^i) \,\bigg|\, \{e_{\tau,h}^i\}_{\tau,i=1}^{L,N}\bigg)
\end{aligned}$$

$$\leq 2\mathcal{N}_1\bigg(\frac{\varepsilon\beta}{5}, \mathcal{F}_{\tilde{\mathcal{W}}}, \{e^i_{\tau,h}\}_{\tau,i=1}^{L,N}\bigg) \exp\bigg(-\frac{\varepsilon^2(1-\varepsilon)\alpha NL}{20(B_S+rB_{\bar{K}})^2(1+\varepsilon)}\bigg) \tag{72}$$

where the first inequality results from the union bound.

Step 3: Bounding the covering number.

In this step, we upper bound the covering number of $\mathcal{F}_{\tilde{\mathcal{W}}}$ by the covering numbers of $\mathbb{B}(r,\bar{\mathcal{H}})$ and $\tilde{\mathcal{W}}$ and conclude the tail probability. We note that

$$\begin{aligned}&\frac{1}{NL}\sum_{\tau,i}\big|f_W(e^i_{\tau,h})-\bar{f}_W(e^i_{\tau,h})\big|\\&\quad\leq 2(B_S+rB_{\bar{K}})\big[B_{\bar{K}}\|f-\bar{f}\|_{\bar{\mathcal{H}}}+rL_KB_k\|W-\bar{W}\|_\infty\big],\end{aligned}$$

where the inequality results from Lemma 33 and the triangle inequality. Thus, we have that

$$\mathcal{N}_1\big(\delta,\mathcal{F}_{\tilde{\mathcal{W}}},\{e^i_{\tau,h}\}_{\tau,i=1}^{L,N}\big)\leq \mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{\delta}{4(B_S+rB_{\bar{K}})B_{\bar{K}}},\mathbb{B}(r,\bar{\mathcal{H}})\bigg)\cdot\mathcal{N}_\infty\bigg(\frac{\delta}{4(B_S+rB_{\bar{K}})rL_KB_k},\tilde{\mathcal{W}}\bigg) \tag{73}$$

for any $\{e^i_{\tau,h}\}_{\tau,i=1}^{L,N}$. Combining the inequalities (66), (67), (68), (70), (71), and (72), we have that for $NL\geq 32(B_S+rB_{\bar{K}})^2/((\alpha+\beta)\varepsilon^2)$

$$\begin{aligned}&\mathbb{P}\bigg(\exists f_W\in\mathcal{F}_{\tilde{\mathcal{W}}},\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]-\frac{1}{NL}\sum_{\tau,i}f_W(e^i_{\tau,h})\\&\qquad\geq\varepsilon\bigg(\alpha+\beta+\frac{1}{NL}\sum_{\tau,i}\mathbb{E}_{\rho^i_{\tau,h}}\big[f_W(e^i_{\tau,h})\big]\bigg)\bigg)\\&\quad\leq\frac{64}{7}\mathbb{E}\bigg[\mathcal{N}_1\bigg(\frac{\varepsilon(\alpha+\beta)}{10(B_S+rB_{\bar{K}})^2},\mathcal{F}_{\tilde{\mathcal{W}}},\{e^i_{\tau,h}\}_{\tau,i=1}^{L,N}\bigg)\bigg]\exp\bigg(-\frac{3\varepsilon^2(\alpha+\beta)NL}{40(B_S+rB_{\bar{K}})^4}\bigg)\\&\qquad+\frac{32}{7}\mathbb{E}\bigg[\mathcal{N}_1\bigg(\frac{\varepsilon\beta}{5},\mathcal{F}_{\tilde{\mathcal{W}}},\{e^i_{\tau,h}\}_{\tau,i=1}^{L,N}\bigg)\bigg]\exp\bigg(-\frac{\varepsilon^2(1-\varepsilon)\alpha NL}{20(B_S+rB_{\bar{K}})^2(1+\varepsilon)}\bigg)\\&\quad\leq 14\mathbb{E}\bigg[\mathcal{N}_1\bigg(\frac{\varepsilon\beta}{10(B_S+rB_{\bar{K}})^2},\mathcal{F}_{\tilde{\mathcal{W}}},\{e^i_{\tau,h}\}_{\tau,i=1}^{L,N}\bigg)\bigg]\exp\bigg(-\frac{\varepsilon^2(1-\varepsilon)\alpha NL}{20(B_S+rB_{\bar{K}})^4(1+\varepsilon)}\bigg)\\&\quad\leq 14\mathcal{N}_{\bar{\mathcal{H}}}\bigg(\frac{\varepsilon\beta}{40(B_S+rB_{\bar{K}})^3B_{\bar{K}}},\mathbb{B}(r,\bar{\mathcal{H}})\bigg)\cdot\mathcal{N}_\infty\bigg(\frac{\varepsilon\beta}{40(B_S+rB_{\bar{K}})^3rL_KB_k},\tilde{\mathcal{W}}\bigg)\\&\qquad\cdot\exp\bigg(-\frac{\varepsilon^2(1-\varepsilon)\alpha NL}{20(B_S+rB_{\bar{K}})^4(1+\varepsilon)}\bigg),\end{aligned}$$

where the last inequality results from inequality (73). For $NL\leq 32(B_S+rB_{\bar{K}})^2/((\alpha+\beta)\varepsilon^2)$, we have that

$$\exp\bigg(-\frac{\varepsilon^2(1-\varepsilon)\alpha NL}{20(B_S+rB_{\bar{K}})^4(1+\varepsilon)}\bigg)\geq\exp\bigg(-\frac{32(1-\varepsilon)\alpha}{20(B_S+rB_{\bar{K}})^2(1+\varepsilon)(\alpha+\beta)}\bigg)\geq\exp\bigg(-\frac{32}{80}\bigg)\geq\frac{1}{14}.$$

Thus, we conclude the proof of Proposition 20. ■

Q.1.3 Proof of Proposition 26

Proof [Proof of Proposition 26] The proof of Proposition 26 mainly follows the pipeline of the proof of Györfi et al. (2002, Theorem 11.6). However, the random variables in our problem are not identically distributed, which requires additional techniques to control the tail probabilities. Our proof involves two steps:

- Symmetrization by a ghost sample.
- Additional randomization by random signs

Step 1: Symmetrization by a ghost sample.

We draw ghost samples $\tilde{Z}_1^n = (\tilde{Z}_1, \cdots, \tilde{Z}_n)$ that are independent of and identically distributed as $Z_1^n = (Z_1, \cdots, Z_n)$. Then we have that

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{g\in\mathcal{G}} \frac{\frac{1}{n}\sum_{i=1}^n g(\tilde{Z}_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]}{\alpha + \frac{1}{n}\sum_{i=1}^n g(\tilde{Z}_i) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]} > \beta\bigg) \\
&\quad \le \frac{\mathbb{E}\Big[\Big(\sum_{i=1}^n g(\tilde{Z}_i) - \mathbb{E}_{\rho_i}\big[g(Z)\big]\Big)^2\Big]}{n^2\beta^2\Big(\alpha + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]\Big)^2} \\
&\quad \le \frac{\sum_{i=1}^n \Big(B - \mathbb{E}_{\rho_i}\big[g(Z)\big]\Big)\mathbb{E}_{\rho_i}\big[g(Z)\big]}{n^2\beta^2\Big(\alpha + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]\Big)^2}, \qquad (74)
\end{aligned}$$

where the first inequality results from Chebyshev inequality, and the last inequality results from that $g : \mathcal{X} \to [0, B]$. For two constants $a, b > 0$ and variables $0 \le x_i \le b$ for $i \in [n]$, some basic calculus calculations show that

$$f(x_1, \cdots, x_n) = \frac{\sum_{i=1}^n (b - x_i)x_i}{\big(a + \frac{1}{n}\sum_{i=1}^n x_i\big)} \le \frac{nb}{2a}.$$

Thus, inequality (74) shows that

$$\mathbb{P}\bigg(\sup_{g\in\mathcal{G}} \frac{\frac{1}{n}\sum_{i=1}^n g(\tilde{Z}_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]}{\alpha + \frac{1}{n}\sum_{i=1}^n g(\tilde{Z}_i) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]} > \beta\bigg) \le \frac{B}{2\beta^2\alpha n}.$$

We take $\beta = \varepsilon/4$. If $n \ge 16B/(\varepsilon^2\alpha)$, such probability is upper bounded by $1/2$. Then we have that

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{g\in\mathcal{G}} \frac{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]}{\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\rho_i}\big[g(Z)\big]} > \varepsilon\bigg) \\
&\quad \le 2\mathbb{P}\bigg(\exists g \in \mathcal{G}, \frac{1}{n}\sum_{i=1}^n \big(g(Z_i) - g(\tilde{Z}_i)\big) \ge \frac{3\varepsilon}{8}\bigg(2\alpha + \frac{1}{n}\sum_{i=1}^n \big(g(Z_i) + g(\tilde{Z}_i)\big)\bigg)\bigg), \qquad (75)
\end{aligned}$$

where the inequality results from the conditional probability trick. The detailed procedure can be found in Györfi et al. (2002, Theorem 11.6).

Step 2: Additional randomization by random signs.

Let $\{U_i\}_{i=1}^n$ be independent and uniformly distributed random variables on $\{+1, 1\}$ that are independent of Z_1^n and $\tilde{Z}_1^n$. Then we have that

$$\begin{aligned}
&\mathbb{P}\bigg(\exists g \in \mathcal{G}, \frac{1}{n}\sum_{i=1}^n \big(g(Z_i) - g(\tilde{Z}_i)\big) \geq \frac{3\varepsilon}{8}\bigg(2\alpha + \frac{1}{n}\sum_{i=1}^n \big(g(Z_i) + g(\tilde{Z}_i)\big)\bigg)\bigg) \\
&\quad \leq 2\mathbb{E}\bigg[\mathbb{P}\bigg(\exists g \in \mathcal{G}, \frac{1}{n}\sum_{i=1}^n U_i g(Z_i) \geq \frac{3\varepsilon}{8}\bigg(\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i)\bigg) \,\bigg|\, Z_1^n = z_1^n\bigg)\bigg], \qquad (76)
\end{aligned}$$

where the inequality results from the union bound. Let $\delta > 0$, $\mathcal{G}_\delta$ be a L_1 $\delta-$cover of $\mathcal{G}$ on z_1^n. Then for any $g \in \mathcal{G}$, there exists $\bar{g} \in \mathcal{G}_\delta$ such that $\sum_{i=1}^n |g(z_i) - \bar{g}(z_i)|/n \leq \delta$. Thus, we have that

$$\begin{aligned}
&\mathbb{P}\bigg(\exists g \in \mathcal{G}, \frac{1}{n}\sum_{i=1}^n U_i g(Z_i) \geq \frac{3\varepsilon}{8}\bigg(\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i)\bigg) \,\bigg|\, Z_1^n = z_1^n\bigg) \\
&\quad \leq \mathbb{P}\bigg(\exists g \in \mathcal{G}_\delta, \delta + \frac{1}{n}\sum_{i=1}^n U_i g(Z_i) \geq \frac{3\varepsilon}{8}\bigg(\alpha - \delta + \frac{1}{n}\sum_{i=1}^n g(Z_i)\bigg) \,\bigg|\, Z_1^n = z_1^n\bigg) \\
&\quad \leq |\mathcal{G}_\delta| \max_{g \in \mathcal{G}_\delta} \mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n U_i g(Z_i) \geq \frac{3\varepsilon\alpha}{8} - \frac{3\varepsilon\delta}{8} - \delta + \frac{3\varepsilon}{8}\frac{1}{n}\sum_{i=1}^n g(Z_i) \,\bigg|\, Z_1^n = z_1^n\bigg),
\end{aligned}$$

where the last inequality follows from the union bound. Take $\delta = \varepsilon\alpha/5$, then we have

$$\frac{3\varepsilon\alpha}{8} - \frac{3\varepsilon\delta}{8} - \delta \geq \frac{\varepsilon\alpha}{10}.$$

Thus, we can control the tail probability as

$$\begin{aligned}
&\mathbb{P}\bigg(\exists g \in \mathcal{G}, \frac{1}{n}\sum_{i=1}^n U_i g(Z_i) \geq \frac{3\varepsilon}{8}\bigg(\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i)\bigg) \,\bigg|\, Z_1^n = z_1^n\bigg) \\
&\quad \leq \mathcal{N}_1\bigg(\frac{\varepsilon\alpha}{5}, \mathcal{G}, z_1^n\bigg) \max_{g \in \mathcal{G}_{\frac{\varepsilon\alpha}{5}}} \mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n U_i g(Z_i) \geq \frac{\varepsilon\alpha}{10} + \frac{3\varepsilon}{8}\frac{1}{n}\sum_{i=1}^n g(Z_i) \,\bigg|\, Z_1^n = z_1^n\bigg) \\
&\quad \leq \mathcal{N}_1\bigg(\frac{\varepsilon\alpha}{5}, \mathcal{G}, z_1^n\bigg) \exp\bigg(-\frac{9\varepsilon^2}{128B}\frac{\big(\frac{4}{15}n\alpha + \sum_{i=1}^n g(z_i)\big)^2}{\sum_{i=1}^n g(z_i)}\bigg) \\
&\quad \leq \mathcal{N}_1\bigg(\frac{\varepsilon\alpha}{5}, \mathcal{G}, z_1^n\bigg) \exp\bigg(-\frac{3\alpha\varepsilon^2 n}{40B}\bigg), \qquad (77)
\end{aligned}$$

where the second inequality results from the Hoeffding's inequality, and the last inequality results from that $(a + y)^2/y \geq 4a$ for any $a, y > 0$. Combining the inequalities (75), (76) and (77), we conclude the proof of Proposition 26. ■

Q.2 Propositions and Lemmas for Optimization

Q.2.1 Proof of Proposition 13

Proof [Proof of Proposition 13] From the definition of $R(\cdot)$ and $\mathrm{KL}(\cdot\|\cdot)$, we have that

$$\nabla_p R(p) = 1 + \log p \quad \nabla_p \mathrm{KL}(p\|q) = 1 + \log\frac{p}{q}.$$

Then the first-order optimal condition of Eqn. (16) is that for any $p \in \Delta(\mathcal{A})$

$$\Big\langle \eta_{t+1}\hat{Q}_h^{\lambda,\alpha}(s,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}) - \lambda\eta_{t+1}\log\hat{\pi}_{t+1,h}^\alpha(\cdot\,|\,s) - \log\frac{\hat{\pi}_{t+1,h}^\alpha(\cdot\,|\,s)}{\pi_{t,h}^\alpha(\cdot\,|\,s)}, p - \hat{\pi}_{t+1,h}^\alpha(\cdot\,|\,s)\Big\rangle \le 0.$$

Note that

$$\begin{aligned}
\mathrm{KL}(p_1\|p_2) &= \mathrm{KL}(p_3\|p_2) + \big\langle \nabla_{p_3}\mathrm{KL}(p_3\|p_2), p_1 - p_3\big\rangle + \mathrm{KL}(p_1\|p_3) \\
\mathrm{KL}(p_1\|p_2) &= R(p_1) - R(p_2) + \big\langle \nabla R(p_2), p_2 - p_1\big\rangle.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\eta_{t+1}\langle \hat{Q}_h^{\lambda,\alpha}(s,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}), p - \hat{\pi}_{t+1,h}^\alpha(\cdot|s)\rangle + \lambda\eta_{t+1}\Big[R\big(\hat{\pi}_{t+1,h}^\alpha(\cdot|s)\big) - R(p)\Big] + \mathrm{KL}\big(\hat{\pi}_{t+1,h}^\alpha(\cdot|s)\|\pi_{t,h}^\alpha(\cdot|s)\big) \\
&\quad \le \mathrm{KL}\big(p\|\pi_{t,h}^\alpha(\cdot\,|\,s)\big) - (1+\lambda\eta_{t+1})\mathrm{KL}\big(p\|\hat{\pi}_{t+1,h}^\alpha(\cdot\,|\,s)\big).
\end{aligned}$$

Thus, we conclude the proof of Proposition 13. ■

Q.2.2 Proof of Proposition 14

Proof [Proof of Proposition 14] In the following, we upper bound these four terms separately. For term (I), we have that

$$\begin{aligned}
\text{(I)} &\le \eta_{t+1}\Big|\big\langle Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\bar{\mu}_t^{\mathcal{I}},W^*), p - \pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle - \big\langle \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}), p - \hat{\pi}_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle\Big| \\
&\le \eta_{t+1}\Big|\big\langle Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\bar{\mu}_t^{\mathcal{I}},W^*) - \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}), p - \pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle\Big| \\
&\qquad + \eta_{t+1}\Big|\big\langle \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W}), \hat{\pi}_{t+1,h}^\alpha(\cdot\,|\,s_h) - \pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big\rangle\Big| \\
&\le 2\eta_{t+1}\big\|Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\bar{\mu}_t^{\mathcal{I}},W^*) - Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*)\big\|_\infty \\
&\qquad + 2\eta_{t+1}\big\|Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},W^*) - \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^\alpha,\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})\big\|_\infty \\
&\qquad + 2\eta_{t+1}H(1+\lambda\log|\mathcal{A}|)\beta_{t+1},
\end{aligned} \tag{78}$$

where the second inequality results from the triangle inequality, and the last inequality results from the Hölder inequality and the triangle inequality. To bound the second term in the right-hand side of inequality (78), we state the proposition

Proposition 27 *Under Assumption 1, for any policy $\pi^{\mathcal{I}}$ and two distribution flows $\mu^{\mathcal{I}}$ and $\tilde{\mu}^{\mathcal{I}}$, we have that*

$$\big|Q_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},W^*) - Q_h^{\lambda,\alpha}(s,a,\pi^\alpha,\tilde{\mu}^{\mathcal{I}},W^*)\big| \le \big[L_r + H(1+\lambda\log|\mathcal{A}|)L_P\big]\sum_{m=h}^H\int_0^1\|\mu_m^\beta - \tilde{\mu}_m^\beta\|_1\mathrm{d}\beta,$$

$$\big|V_h^{\lambda,\alpha}(s,\pi^\alpha,\mu^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s,\pi^\alpha,\tilde{\mu}^{\mathcal{I}},W^*)\big| \le \big[L_r + H(1+\lambda\log|\mathcal{A}|)L_P\big]\sum_{m=h}^H\int_0^1\|\mu_m^\beta - \tilde{\mu}_m^\beta\|_1\mathrm{d}\beta$$

for all $\alpha \in \mathcal{I}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $h \in [H]$.

Proof [Proof of Proposition 27] See Appendix Q.2.8. ∎

Thus, we have that

$$
\begin{aligned}
\text{(I)} \leq 2\eta_{t+1}\big[L_r + H(1+\lambda\log|\mathcal{A}|)L_P\big]\sum_{m=1}^{H}\int_0^1 \|\bar{\mu}_{t,m}^{\beta} - \hat{\bar{\mu}}_{t,m}^{\beta}\|_1 \mathrm{d}\beta \\
+ 2\eta_{t+1}\big\|Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^{\alpha},\hat{\bar{\mu}}_t^{\mathcal{I}},W^*) - \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^{\alpha},\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})\big\|_\infty + 2\eta_{t+1}H(1+\lambda\log|\mathcal{A}|)\beta_{t+1}.
\end{aligned}
$$

Define $\alpha_{m,t} = \alpha'_m\prod_{k=m+1}^{t-1}(1-\alpha'_k)$ for $m\in[t]$, where $\alpha'_m = \alpha_m$ for $m\geq 2$ and $\alpha'_1 = 1$ (since $\hat{\bar{\mu}}_1^{\mathcal{I}} = \hat{\mu}_1^{\mathcal{I}}$). Then it satisfies that $\sum_{m=1}^{t-1}\alpha_{m,t} = 1$, and that

$$
\bar{\mu}_t^{\mathcal{I}} = \sum_{m=1}^{t-1}\alpha_{m,t}\cdot\mu_m^{\mathcal{I}}, \text{ and } \quad \hat{\bar{\mu}}_t^{\mathcal{I}} = \sum_{m=1}^{t-1}\alpha_{m,t}\cdot\hat{\mu}_m^{\mathcal{I}}. \tag{79}
$$

Then we have that

$$
\sum_{m=1}^{H}\int_0^1 \|\bar{\mu}_{t,m}^{\beta} - \hat{\bar{\mu}}_{t,m}^{\beta}\|_1\mathrm{d}\beta = d(\bar{\mu}_t^{\mathcal{I}},\hat{\bar{\mu}}_t^{\mathcal{I}}) \leq \sum_{m=1}^{t-1}\alpha_{m,t-1}d(\hat{\mu}_m^{\mathcal{I}},\mu_m^{\mathcal{I}}) \leq \varepsilon_\mu,
$$

where the inequality results from the triangle inequality. Thus, we have

$$
\begin{aligned}
\text{(I)} \leq 2\eta_{t+1}\big\|Q_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^{\alpha},\hat{\bar{\mu}}_t^{\mathcal{I}},W^*) - \hat{Q}_h^{\lambda,\alpha}(s_h,\cdot,\pi_t^{\alpha},\hat{\bar{\mu}}_t^{\mathcal{I}},\hat{W})\big\|_\infty \\
+ 2\eta_{t+1}\big[L_r + H(1+\lambda\log|\mathcal{A}|)L_P\big]\varepsilon_\mu + 2\eta_{t+1}H(1+\lambda\log|\mathcal{A}|)\beta_{t+1}.
\end{aligned}
$$

For term (II), Lemma 31 shows that (II) ≤ 0.

For term (III), we have that

$$
\begin{aligned}
\text{(III)} &= R\big(\pi_{t+1,h}^{\alpha}(\cdot\,|\,s_h)\big) - R\big(\hat{\pi}_{t+1,h}^{\alpha}(\cdot\,|\,s_h)\big) + \sum_{a\in\mathcal{A}}\big(\pi_{t+1,h}^{\alpha}(a\,|\,s_h) - \hat{\pi}_{t+1,h}^{\alpha}(a\,|\,s_h)\big)\log\frac{1}{\pi_{t,h}^{\alpha}(a\,|\,s_h)} \\
&\leq \sum_{a\in\mathcal{A}}\big|\pi_{t+1,h}^{\alpha}(a\,|\,s_h) - \hat{\pi}_{t+1,h}^{\alpha}(a\,|\,s_h)\big|\log\frac{|\mathcal{A}|}{\beta_t} \\
&\leq 2\beta_{t+1}\log\frac{|\mathcal{A}|}{\beta_t},
\end{aligned}
$$

where the last inequality results from the definition of $\pi_{t+1,h}^{\alpha}$ and $\hat{\pi}_{t+1,h}^{\alpha}$.

For term (IV), Lemma 36 shows that for $\beta_{t+1}\leq 1/2$, we have that (IV) $\leq 2(1+\lambda\eta_{t+1})\beta_{t+1}$.

Summing these four terms, we conclude the proof of the proposition. ∎

Q.2.3 Proof of Proposition 15

Proof [Proof of Proposition 15] Our proof involves two steps:

- Proof $\mathbb{E}_{\pi^*}[V_h^{\lambda}(s_h,\pi^*) - V_h^{\lambda}(s_h,\pi)] \geq \gamma^*\mathbb{E}_{\pi^*}[V_{h+1}^{\lambda}(s_{h+1},\pi^*) - V_{h+1}^{\lambda}(s_{h+1},\pi)]$ for all $h\in[H]$, where $\gamma^*>0$ is a constant.
- Proof the desired result from Step 1.

Step 1: Proof $\mathbb{E}_{\pi^*}[V_h^{\lambda}(s_h,\pi^*)-V_h^{\lambda}(s_h,\pi)] \geq \gamma^*\mathbb{E}_{\pi^*}[V_{h+1}^{\lambda}(s_{h+1},\pi^*)-V_{h+1}^{\lambda}(s_{h+1},\pi)]$ **for all** $h\in[H]$**.**

If $\pi_t=\pi_t^*$ for all $t\geq h+1$, then the result trivially holds. In the following, we assume that $\pi_t\neq\pi_t^*$ for some $t\geq h+1$. This implies that

$$\mathbb{E}_{\pi^*}[V_{h+1}^{\lambda}(s_{h+1},\pi^*)-V_{h+1}^{\lambda}(s_{h+1},\pi)]>0.$$

For ease of notation, we define that

$$\begin{aligned}
y(s,a)&=r_h(s,a)+\int_{\mathcal{S}}P_h(s'\,|\,s,a)V_{h+1}(s',\pi)\mathrm{d}s'\\
y^*(s,a)&=r_h(s,a)+\int_{\mathcal{S}}P_h(s'\,|\,s,a)V_{h+1}(s',\pi^*)\mathrm{d}s'.
\end{aligned}$$

Then we expand these two differences between value functions as

$$\begin{aligned}
&\mathbb{E}_{\pi^*}\big[V_h^{\lambda}(s_h,\pi^*)-V_h^{\lambda}(s_h,\pi)\big]\\
&\quad=\mathbb{E}_{\pi^*}\Big[\langle y(s_h,\cdot),\pi_h^*(\cdot|s_h)-\pi_h(\cdot|s_h)\rangle+\lambda\Big[R\big(\pi_h(\cdot|s_h)\big)-R\big(\pi_h^*(\cdot|s_h)\big)\Big]+\langle y^*(s_h,\cdot)-y(s_h,\cdot),\pi_h^*(\cdot\,|\,s_h)\rangle\Big]\\
&\mathbb{E}_{\pi^*}\big[V_{h+1}^{\lambda}(s_{h+1},\pi^*)-V_{h+1}^{\lambda}(s_{h+1},\pi)\big]\\
&\quad=\mathbb{E}_{\pi^*}\big[\langle y^*(s_h,\cdot)-y(s_h,\cdot),\pi_h^*(\cdot\,|\,s_h)\rangle\big],
\end{aligned}$$

where $R(p)=\langle p,\log p\rangle$. In the following, we will prove that for any $s\in\mathcal{S}$

$$\begin{aligned}
&\langle y(s,\cdot),\pi_h^*(\cdot\,|\,s)-\pi_h(\cdot\,|\,s)\rangle+\lambda\Big[R\big(\pi_h(\cdot\,|\,s)\big)-R\big(\pi_h^*(\cdot\,|\,s)\big)\Big]+\langle y^*(s,\cdot)-y(s,\cdot),\pi_h^*(\cdot\,|\,s)\rangle\\
&\quad\geq\gamma^*\langle y^*(s,\cdot)-y(s,\cdot),\pi_h^*(\cdot\,|\,s)\rangle, \qquad (80)
\end{aligned}$$

and our desired result immediately follows from taking expectation on the both sides of inequality (80). For ease of notation, we define $p^*=\pi_h^*(\cdot\,|\,s)$. From the definition of the optimal policy, we have that

$$p^*=\operatorname*{argmax}_{q\in\Delta(\mathcal{A})}\langle q,y^*(s,\cdot)\rangle-\lambda R(q),\quad p=\operatorname*{argmax}_{q\in\Delta(\mathcal{A})}\langle q,y(s,\cdot)\rangle-\lambda R(q).$$

These distributions admit closed-form expressions $p^*(a)=\exp(y^*(s,a)/\lambda)/Z^*(s)$ and $p(a)=\exp(y(s,a)/\lambda)/Z(s)$, where $Z^*(s)=\sum_a\exp(y^*(s,a)/\lambda)$ and $Z(s)=\sum_a\exp(y(s,a)/\lambda)$. To prove inequality (80), it suffices to prove that

$$\langle y(s,\cdot),p^*-p\rangle+\lambda\big[R(p)-R(p^*)\big]\geq(\gamma^*-1)\langle y^*(s,\cdot)-y(s,\cdot),p^*\rangle. \qquad (81)$$

The left-hand side the inequality (81) is

$$\langle y(s,\cdot),p^*-p\rangle+\lambda\big[R(p)-R(p^*)\big]=\langle\lambda\log p,p^*-p\rangle+\lambda\big[R(p)-R(p^*)\big]=-\lambda\Big\langle p^*,\log\frac{p^*}{p}\Big\rangle, \qquad (82)$$

where the first equality results from the closed-form expression of p, and the second inequality results from the definition of $R(\cdot)$. We further expand this term as

$$-\lambda\Big\langle p^*,\log\frac{p^*}{p}\Big\rangle=-\lambda\log\frac{Z(s)}{Z^*(s)}-\Big\langle\frac{\exp\big(y^*(s,\cdot)/\lambda\big)}{Z^*(s)},y^*(s,\cdot)-y(s,\cdot)\Big\rangle, \qquad (83)$$

where the equalitys result from the closed-form expressions of p and p^*. The right-hand side of inequality (81) is

$$(\gamma^* - 1)\langle y^*(s,\cdot) - y(s,\cdot), p^*\rangle = (\gamma^* - 1)\lambda\bigg(\bigg\langle \log\frac{p^*}{p}, p^*\bigg\rangle + \log\frac{Z^*(s)}{Z(s)}\bigg), \tag{84}$$

where the equalitys result from the closed-form expressions of p and p^*. Combining Eqn. (82), (83), and (84), we have

$$\begin{aligned}&\langle y(s,\cdot), p^* - p\rangle + \lambda\big[R(p) - R(p^*)\big] \geq (\gamma^* - 1)\langle y^*(s,\cdot) - y(s,\cdot), p^*\rangle\\ \Leftrightarrow\ &\frac{\gamma^*}{\lambda}\bigg\langle \exp\bigg(\frac{y^*(s,\cdot)}{\lambda}\bigg), y^*(s,\cdot) - y(s,\cdot)\bigg\rangle \leq Z^*(s)\log\frac{Z^*(s)}{Z(s)}.\end{aligned} \tag{85}$$

In the following, we prove inequality (85). The right-hand side of (85) can be lower-bounded as

$$\begin{aligned}Z^*(s)\log\frac{Z^*(s)}{Z(s)} &\geq \frac{\log B}{B-1}\sum_{a\in\mathcal{A}}\exp\big(y^*(s,a)/\lambda\big)\cdot\bigg[\frac{\sum_{a\in\mathcal{A}}\exp\big(y^*(s,a)/\lambda\big)}{\sum_{a\in\mathcal{A}}\exp\big(y(s,a)/\lambda\big)} - 1\bigg]\\ &\geq \frac{\log B}{(B-1)\lambda}\cdot\sum_{a\in\mathcal{A}}\exp\big(y(s,a)/\lambda\big)\big(y^*(s,a) - y(s,a)\big),\end{aligned} \tag{86}$$

where $B = \exp(H(1 + \lambda\log|\mathcal{A}|)/\lambda)$, the first inequality results from that $\log B/(B-1)\cdot(x - 1) \leq \log x$ for $x \in [1, B]$ and the facts that $y^*(s,a) \geq y(s,a)$ and $|y^*(s,a)| \leq H(1 + \lambda\log|\mathcal{A}|)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and the second inequality results from that $\exp(x) - 1 \geq x$ and that $y^*(s,a) \geq y(s,a)$. The left-hand side of inequality (85) can be upper bounded as

$$\frac{\gamma^*}{\lambda}\bigg\langle \exp\bigg(\frac{y^*(s,\cdot)}{\lambda}\bigg), y^*(s,\cdot) - y(s,\cdot)\bigg\rangle \leq \frac{\gamma^*}{\lambda}\cdot B\cdot\sum_{a\in\mathcal{A}}\exp\big(y(s,a)/\lambda\big)\big(y^*(s,a) - y(s,a)\big), \tag{87}$$

where the inequality results from that $y^*(s,a)/y(s,a) \leq B$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Combining inequalities (86) and (87), we prove inequality (85) given

$$0 < \gamma^* \leq \frac{\log B}{B(B-1)}.$$

Step 2: Proof the desired result from Step 1.
We define that

$$D_h = \mathbb{E}_{\pi^*}[V_h^\lambda(s_h, \pi^*) - V_h^\lambda(s_h, \pi)]$$

for $h \in [H]$. Then Step 1 shows that $D_h \geq \gamma^* D_{h+1}$ for all $h \in [H]$. Thus, we have

$$\frac{\mathbb{E}_{\pi^*}\big[V_1^\lambda(s_1,\pi^*) - V_1^\lambda(s_1,\pi)\big]}{\mathbb{E}_{\pi^*}\Big[\sum_{h=2}^H V_h^\lambda(s_h,\pi^*) - V_h^\lambda(s_h,\pi)\Big]} = \frac{D_1}{\sum_{h=2}^H D_h} = \frac{1}{\sum_{h=2}^H D_h/D_1}.$$

For each term, we have that

$$\frac{D_h}{D_1} = \frac{D_h}{D_{h-1}} \cdots \frac{D_2}{D_1} \le \gamma^{*(2-h)}.$$

Thus, we have that

$$\frac{\mathbb{E}_{\pi^*}\left[V_1^\lambda(s_1,\pi^*) - V_1^\lambda(s_1,\pi)\right]}{\mathbb{E}_{\pi^*}\left[\sum_{h=2}^H V_h^\lambda(s_h,\pi^*) - V_h^\lambda(s_h,\pi)\right]} \ge \frac{(1-\gamma^*)\gamma^{*(H-2)}}{1-\gamma^{*(H-1)}} = \beta^*.$$

The proof of Proposition 15 is complete. ■

Q.2.4 Proof of Proposition 16

Proof [Proof of Proposition 16]
We first write

$$\begin{aligned}
\Delta_{t+1}^\alpha &= X_{t+1}^\alpha - \mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\left[\sum_{h=1}^H V_h^{\lambda,\alpha}(s_h,\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s_h,\pi_{t+1}^\alpha,\bar{\mu}_t^{\mathcal{I}},W^*)\right] \\
&\quad - \frac{1}{\eta\theta^*}\mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\left[\sum_{h=1}^H \mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s_h)\|\pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big)\right] \\
&= \text{(V)} + \text{(VI)} + \text{(VII)} + \text{(VIII)} + \text{(IX)}.
\end{aligned} \tag{88}$$

Term (V) is the error that measures the difference between the action-value function induced by the optimal policies of $\bar{\mu}_{t+1}^{\mathcal{I}}$ and $\bar{\mu}_t^{\mathcal{I}}$, which is defined as

$$\text{(V)} = \mathbb{E}_{\bar{\pi}_{t+1}^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}}}\left[\sum_{h=1}^H V_h^{\lambda,\alpha}(s_h,\bar{\pi}_{t+1}^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s_h,\bar{\pi}_t^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}},W^*)\right].$$

To upper bound the term (V), we note that the optimal policies $\bar{\mu}_{t+1}^{\mathcal{I}}$ and $\bar{\mu}_t^{\mathcal{I}}$ satisfy the following property.

Proposition 28 *For a λ-regularized finite-horizon MDP $(\mathcal{S},\mathcal{A},H,\{r_h\}_{h=1}^H,\{P_h\}_{h=1}^H)$ with $r_h \in [0,1]$ for all $h \in [H]$, we denote the optimal policy as $\pi^* = \{\pi_h^*\}_{h=1}^H$. Then we have that for any $s \in \mathcal{S}$, and $h \in [H]$*

$$\min_{a\in\mathcal{A}} \pi_h^*(a\,|\,s) \ge \frac{1}{1+|\mathcal{A}|\exp\big((H-h+1)(1+\lambda\log|\mathcal{A}|)/\lambda\big)}.$$

Proof [Proof of Proposition 28] See Appendix Q.2.5 ■

Then we have that

$$\begin{aligned}
&\big|V_h^{\lambda,\alpha}(s_h,\bar{\pi}_{t+1}^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}},W^*) - V_h^{\lambda,\alpha}(s_h,\bar{\pi}_t^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}},W^*)\big| \\
&\quad \le \big(H(1+\lambda\log|\mathcal{A}|)+\lambda L_R\big)\mathbb{E}_{\bar{\pi}_{t+1}^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}}}\left[\sum_{m=h}^H \big\|\bar{\pi}_{t+1,m}^{*,\alpha}(\cdot\,|\,s_m) - \bar{\pi}_{t,m}^{*,\alpha}(\cdot\,|\,s_m)\big\|_1\right],
\end{aligned}$$

where $L_R = \log(1+|\mathcal{A}|\exp\big(H(1+\lambda\log|\mathcal{A}|)/\lambda\big)$, the inequality results from the performance difference lemma, Lemma 37, proposition 28 and Lemma 38. Thus, we have that

$$\text{(V)} \le H\big(H(1+\lambda\log|\mathcal{A}|)+\lambda L_R\big)\mathbb{E}_{\bar{\pi}^{*,\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_{t+1}}\Bigg[\sum_{m=1}^{H}\big\|\bar{\pi}^{*,\alpha}_{t+1,m}(\cdot\,|\,s_m)-\bar{\pi}^{*,\alpha}_{t,m}(\cdot\,|\,s_m)\big\|_1\Bigg]. \tag{89}$$

Term (VI) is the error that measures the difference between the distribution of states induced by optimal policies of $\bar{\mu}^{\mathcal{I}}_{t+1}$ and $\bar{\mu}^{\mathcal{I}}_{t}$, which is defined as

$$\begin{aligned}\text{(VI)} &= \mathbb{E}_{\bar{\pi}^{*,\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_{t+1}} - \mathbb{E}_{\bar{\pi}^{*,\alpha}_{t},\bar{\mu}^{\mathcal{I}}_{t+1}}\Bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h,\bar{\pi}^{*,\alpha}_t,\bar{\mu}^{\mathcal{I}}_{t+1},W^*)-V_h^{\lambda,\alpha}(s_h,\pi^{\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_{t+1},W^*)\Bigg]\\ &\quad+\frac{1}{\eta\theta^*}\Big(\mathbb{E}_{\bar{\pi}^{*,\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_{t+1}}-\mathbb{E}_{\bar{\pi}^{*,\alpha}_{t},\bar{\mu}^{\mathcal{I}}_{t+1}}\Big)\Bigg[\sum_{h=1}^{H}\mathrm{KL}\big(\bar{\pi}^{*,\alpha}_{t,h}(\cdot\,|\,s_h)\|\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big)\Bigg].\end{aligned}$$

Proposition 29 *Given any two policies $\pi^{\mathcal{I}},\tilde{\pi}^{\mathcal{I}}$ and distribution flow $\mu^{\mathcal{I}}$, we define $\mu^{+,\mathcal{I}}=\Gamma_3(\pi^{\mathcal{I}},\mu^{\mathcal{I}},W)$ and $\tilde{\mu}^{+,\mathcal{I}}=\Gamma_3(\tilde{\pi}^{\mathcal{I}},\mu^{\mathcal{I}},W)$ for any graphons $W=\{W_h\}_{h=1}^{H}$. Then we have*

$$\big\|\mu_h^{+,\alpha}-\tilde{\mu}_h^{+,\alpha}\big\|_1 \le \sum_{m=1}^{h-1}\mathbb{E}_{\mu_m^{+,\alpha}}\Big[\big\|\pi^{\alpha}_m(\cdot\,|\,s)-\tilde{\pi}^{\alpha}_m(\cdot\,|\,s)\big\|_1\Big].$$

for all $\alpha\in\mathcal{I}$ and $h\in[H]$.

Proof [Proof of Proposition 29] See Appendix Q.2.6. ■

Proposition 29 shows that

$$\begin{aligned}\text{(VI)} &\le H\Big(H(1+\lambda\log|\mathcal{A}|)+\max_{s\in\mathcal{S},h\in[H]}\mathrm{KL}\big(\bar{\pi}^{*,\alpha}_{t,h}(\cdot\,|\,s)\|\pi^{\alpha}_{t+1,h}(\cdot\,|\,s)\big)\Big)\\ &\quad\cdot\mathbb{E}_{\bar{\pi}^{*,\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_{t+1}}\Bigg[\sum_{m=1}^{H}\big\|\bar{\pi}^{*,\alpha}_{t+1,m}(\cdot\,|\,s_m)-\bar{\pi}^{*,\alpha}_{t,m}(\cdot\,|\,s_m)\big\|_1\Bigg].\end{aligned}$$

Note that

$$\mathrm{KL}\big(\bar{\pi}^{*,\alpha}_{t,h}(\cdot\,|\,s)\|\pi^{\alpha}_{t+1,h}(\cdot\,|\,s)\big)\le\log|\mathcal{A}|+\log\frac{|\mathcal{A}|}{\beta_{t+1}}.$$

Thus, we have

$$\text{(VI)}\le H\Bigg(H\big(1+\lambda\log|\mathcal{A}|\big)+\frac{1}{\eta\theta^*}\log\frac{|\mathcal{A}|^2}{\beta_{t+1}}\Bigg)\cdot\mathbb{E}_{\bar{\pi}^{*,\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_{t+1}}\Bigg[\sum_{m=1}^{H}\big\|\bar{\pi}^{*,\alpha}_{t+1,m}(\cdot\,|\,s_m)-\bar{\pi}^{*,\alpha}_{t,m}(\cdot\,|\,s_m)\big\|_1\Bigg]. \tag{90}$$

Term (VII) is the error that measures the difference between the distribution of states induced by $\bar{\mu}^{\mathcal{I}}_{t}$ on $\bar{\mu}^{\mathcal{I}}_{t+1}$ and $\bar{\mu}^{\mathcal{I}}_{t}$, which is defined as

$$\begin{aligned}\text{(VII)} &= \mathbb{E}_{\bar{\pi}^{*,\alpha}_{t},\bar{\mu}^{\mathcal{I}}_{t+1}} - \mathbb{E}_{\bar{\pi}^{*,\alpha}_{t},\bar{\mu}^{\mathcal{I}}_{t}}\Bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h,\bar{\pi}^{*,\alpha}_t,\bar{\mu}^{\mathcal{I}}_{t+1},W^*)-V_h^{\lambda,\alpha}(s_h,\pi^{\alpha}_{t+1},\bar{\mu}^{\mathcal{I}}_{t+1},W^*)\Bigg]\\ &\quad+\frac{1}{\eta\theta^*}\Big(\mathbb{E}_{\bar{\pi}^{*,\alpha}_{t},\bar{\mu}^{\mathcal{I}}_{t+1}}-\mathbb{E}_{\bar{\pi}^{*,\alpha}_{t},\bar{\mu}^{\mathcal{I}}_{t}}\Big)\Bigg[\sum_{h=1}^{H}\mathrm{KL}\big(\bar{\pi}^{*,\alpha}_{t,h}(\cdot\,|\,s_h)\|\pi^{\alpha}_{t+1,h}(\cdot\,|\,s_h)\big)\Bigg].\end{aligned}$$

Proposition 30 *Given any policy $\pi^{\mathcal{I}}$ and two distribution flows $\mu^{\mathcal{I}}$ and $\tilde{\mu}^{\mathcal{I}}$, we define $\mu^{+,\mathcal{I}} = \Gamma_3(\pi^{\mathcal{I}}, \mu^{\mathcal{I}}, W^*)$ and $\tilde{\mu}^{+,\mathcal{I}} = \Gamma_3(\pi^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}}, W^*)$. Under Assumption 1, we have that*

$$\big\|\mu_h^{+,\alpha} - \tilde{\mu}_h^{+,\alpha}\big\|_1 \leq L_P \sum_{m=1}^{h-1} \int_0^1 \|\mu_m^\beta - \tilde{\mu}_m^\beta\|_1 \mathrm{d}\beta$$

for all $\alpha \in \mathcal{I}$ and $h \in [H]$.

Proof [Proof of Proposition 30] See Appendix Q.2.7. ■

Following the similar arguments in inequality (90), we have that

$$\text{(VII)} \leq H\bigg(H\big(1 + \lambda \log|\mathcal{A}|\big) + \frac{1}{\eta\theta^*} \log \frac{|\mathcal{A}|^2}{\beta_{t+1}}\bigg) L_P \cdot \sum_{m=1}^{H} \int_0^1 \|\bar{\mu}_{t+1,m}^\beta - \bar{\mu}_{t,m}^\beta\|_1 \mathrm{d}\beta. \tag{91}$$

Term (VIII) is the error that measures the difference between the action-value function induced by difference distribution flows $\bar{\mu}_{t+1}^{\mathcal{I}}$ and $\bar{\mu}_t^{\mathcal{I}}$, which is defined as

$$\begin{aligned}
\text{(VIII)} = \mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h, \bar{\pi}_t^{*,\alpha}, \bar{\mu}_{t+1}^{\mathcal{I}}, W^*) - V_h^{\lambda,\alpha}(s_h, \pi_{t+1}^\alpha, \bar{\mu}_{t+1}^{\mathcal{I}}, W^*)\bigg] \\
- \mathbb{E}_{\bar{\pi}_t^{*,\alpha},\bar{\mu}_t^{\mathcal{I}}}\bigg[\sum_{h=1}^{H} V_h^{\lambda,\alpha}(s_h, \bar{\pi}_t^{*,\alpha}, \bar{\mu}_t^{\mathcal{I}}, W^*) - V_h^{\lambda,\alpha}(s_h, \pi_{t+1}^\alpha, \bar{\mu}_t^{\mathcal{I}}, W^*)\bigg]
\end{aligned}$$

From Proposition 27, we have that

$$\text{(VIII)} \leq 2H\big[L_r + H(1 + \lambda \log|\mathcal{A}|)L_P\big] \sum_{m=1}^{H} \int_0^1 \|\mu_m^\beta - \tilde{\mu}_m^\beta\|_1 \mathrm{d}\beta. \tag{92}$$

Term (IX) is the error that measures the difference between the KL divergence related to the optimal policies of $\bar{\mu}_{t+1}^{\mathcal{I}}$ and $\bar{\mu}_t^{\mathcal{I}}$, which is defined as

$$\text{(IX)} = \frac{1}{\eta\theta^*} \mathbb{E}_{\bar{\pi}_{t+1}^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}}}\bigg[\sum_{h=1}^{H} \mathrm{KL}\big(\bar{\pi}_{t+1,h}^{*,\alpha}(\cdot\,|\,s_h)\|\pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big) - \mathrm{KL}\big(\bar{\pi}_{t,h}^{*,\alpha}(\cdot\,|\,s_h)\|\pi_{t+1,h}^\alpha(\cdot\,|\,s_h)\big)\bigg].$$

Lemma 39 and Proposition 28 show that

$$\text{(IX)} \leq \frac{2}{\eta\theta^*} \max\bigg\{\log \frac{|\mathcal{A}|}{\beta_{t+1}}, L_R\bigg\} \mathbb{E}_{\bar{\pi}_{t+1}^{*,\alpha},\bar{\mu}_{t+1}^{\mathcal{I}}}\bigg[\sum_{m=1}^{H} \big\|\bar{\pi}_{t+1,m}^{*,\alpha}(\cdot\,|\,s_m) - \bar{\pi}_{t,m}^{*,\alpha}(\cdot\,|\,s_m)\big\|_1\bigg]. \tag{93}$$

Combining Eqn. (88) and inequalities (89), (90), (91), (92), (93), we conclude the proof of this proposition. ■

Q.2.5 Proof of Proposition 28

Proof [Proof of Proposition 28] We denote the value function of the optimal policy π^* as $V_h^\lambda(s,\pi^*)$ for $h\in[H]$. From the definition of the optimal policy, we have that for any $s\in\mathcal{S}$

$$\pi_h^*(\cdot\,|\,s)=\operatorname*{argmax}_{p\in\Delta(\mathcal{A})}\langle r_h(s,\cdot),p\rangle-\lambda R(p)+\sum_{a\in\mathcal{A}}\int_{\mathcal{S}}p(a)P_h(s'\,|\,s,a)V_{h+1}^\lambda(s',\pi^*)\mathrm{d}s'.$$

Then we have that

$$\pi_h^*(a\,|\,s)\propto\exp\biggl(\frac{1}{\lambda}\Bigl(r_h(s,a)+\int_{\mathcal{S}}P_h(s'\,|\,s,a)V_{h+1}^\lambda(s',\pi^*)\mathrm{d}s'\Bigr)\biggr).$$

The desired result follows from that $V_h^\lambda(s',\pi^*)\le(H-h+1)(1+\lambda\log|\mathcal{A}|)$. Thus, we conclude the proof of Proposition 28. ■

Q.2.6 Proof of Proposition 29

Proof [Proof of Proposition 29] For any $h\in[H-1]$ and $\alpha\in\mathcal{I}$, we have

$$\begin{aligned}
&\bigl\|\mu_{h+1}^{+,\alpha}-\tilde{\mu}_{h+1}^{+,\alpha}\bigr\|_1\\
&\quad\le\sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\int_{\mathcal{S}}\bigl|\mu_h^{+,\alpha}(s)\pi_h^\alpha(a\,|\,s)-\tilde{\mu}_h^{+,\alpha}(s)\tilde{\pi}_h^\alpha(a\,|\,s)\bigr|P_h\bigl(s'\,|\,s,a,z_h^\alpha(\mu_h^{\mathcal{I}},W_h)\bigr)\mathrm{d}s\mathrm{d}s'\\
&\quad\le\sum_a\int_S\mu_h^{+,\alpha}(s)\bigl|\pi_h^\alpha(a\,|\,s)-\tilde{\pi}_h^\alpha(a\,|\,s)\bigr|\mathrm{d}s+\sum_a\int_S\bigl|\mu_h^{+,\alpha}(s)-\tilde{\mu}_h^{+,\alpha}(s)\bigr|\tilde{\pi}_h^\alpha(a\,|\,s)\mathrm{d}s\\
&\quad=\mathbb{E}_{\mu_h^{+,\alpha}}\Bigl[\bigl\|\pi_h^\alpha(\cdot\,|\,s)-\tilde{\pi}_h^\alpha(\cdot\,|\,s)\bigr\|_1\Bigr]+\bigl\|\mu_h^{+,\alpha}-\tilde{\mu}_h^{+,\alpha}\bigr\|_1,
\end{aligned}$$

where the first inequality results from the definition of Γ_3 and the triangle inequality, and the second inequality results from the triangle inequality. Note that $\mu_1^{+,\alpha}=\tilde{\mu}_1^{+,\alpha}=\mu_1^\alpha$. Summing over h, we prove the desired result. This completes the proof of Proposition 29. ■

Q.2.7 Proof of Proposition 30

Proof [Proof of Proposition 30] For any $h\in[H-1]$, we have that

$$\begin{aligned}
&\bigl\|\mu_{h+1}^{+,\alpha}-\tilde{\mu}_{h+1}^{+,\alpha}\bigr\|_1\\
&\quad\le\int_{\mathcal{S}}\biggl|\sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\bigl(\mu_h^{+,\alpha}(s)-\tilde{\mu}_h^{+,\alpha}(s)\bigr)\pi_h^\alpha(a\,|\,s)P_h^*\bigl(s'\,|\,s,a,z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)\bigr)\biggr|\mathrm{d}s'\\
&\qquad+\int_{\mathcal{S}}\biggl|\sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\tilde{\mu}_h^{+,\alpha}(s)\pi_h^\alpha(a\,|\,s)\Bigl(P_h^*\bigl(s'\,|\,s,a,z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)\bigr)-P_h^*\bigl(s'\,|\,s,a,z_h^\alpha(\tilde{\mu}_h^{\mathcal{I}},W_h^*)\bigr)\Bigr)\biggr|\mathrm{d}s'\\
&\quad\le\|\mu_h^{+,\alpha}-\tilde{\mu}_h^{+,\alpha}\|_1+L_P\bigl\|z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)-z_h^\alpha(\tilde{\mu}_h^{\mathcal{I}},W_h^*)\bigr\|_1,
\end{aligned}$$

where the first inequality results from the definition of Γ_3 and triangle inequality, and the second inequality results from Assumption 1. For the right-hand side term, we have that

$$\big\|z_h^\alpha(\mu_h^{\mathcal{I}}, W_h^*) - z_h^\alpha(\tilde{\mu}_h^{\mathcal{I}}, W_h^*)\big\|_1 = \int_{\mathcal{S}} \bigg| \int_0^1 W_h^*(\alpha, \beta)\big(\mu_h^\beta(s) - \tilde{\mu}_h^\beta(s)\big)\mathrm{d}\beta \bigg| \mathrm{d}s \leq \int_0^1 \|\mu_h^\beta - \tilde{\mu}_h^\beta\|_1 \mathrm{d}\beta,$$

where the inequality results from the triangle inequality and that $|W_h^*| \leq 1$. Summing over h, we prove the desired result. Thus, we conclude the proof of Proposition 30. ■

Q.2.8 Proof of Proposition 27

Proof [Proof of Proposition 27] From the definitions of the value function and the action-value function, we have that for any $h \in [H]$

$$Q_h^{\lambda,\alpha}(s, a, \pi^\alpha, \mu^{\mathcal{I}}, W^*) = r_h^*\big(s, a, z_h^\alpha(\mu^{\mathcal{I}}, W^*)\big) + \int_{\mathcal{S}} \Big[\sum_{a' \in \mathcal{A}} \pi_{h+1}^\alpha(a' \,|\, s') Q_{h+1}^{\lambda,\alpha}(s', a', \pi^\alpha, \mu^{\mathcal{I}}, W^*) - \lambda R\big(\pi_{h+1}^\alpha(\cdot \,|\, s')\big) \Big] P_h^*\big(s' \,|\, s, a, z_h^\alpha(\mu^{\mathcal{I}}, W^*)\big) \mathrm{d}s'.$$

Thus, we have that

$$\begin{aligned} &\big|Q_h^{\lambda,\alpha}(s, a, \pi^\alpha, \mu^{\mathcal{I}}, W^*) - Q_h^{\lambda,\alpha}(s, a, \pi^\alpha, \tilde{\mu}^{\mathcal{I}}, W^*)\big| \\ &\quad \leq \big[L_r + H(1 + \lambda \log|\mathcal{A}|)L_P\big] \int_0^1 \|\mu_h^\beta - \tilde{\mu}_h^\beta\|_1 \mathrm{d}\beta \\ &\qquad + \sum_{a' \in \mathcal{A}} \int_{\mathcal{S}} \big|Q_{h+1}^{\lambda,\alpha}(s', a', \pi^{\mathcal{I}}, \mu^{\mathcal{I}}, W^*) - Q_{h+1}^{\lambda,\alpha}(s', a', \pi^{\mathcal{I}}, \tilde{\mu}^{\mathcal{I}}, W^*)\big| \pi_{h+1}^\alpha(a'|s') P_h^*\big(s'|s, a, z_h^\alpha(\mu^{\mathcal{I}}, W^*)\big) \mathrm{d}s', \end{aligned}$$

where the inequality results from the triangle inequality and Assumption 1. By induction, it is easy to prove that

$$\big|Q_h^{\lambda,\alpha}(s, a, \pi^\alpha, \mu^{\mathcal{I}}, W^*) - Q_h^{\lambda,\alpha}(s, a, \pi^\alpha, \tilde{\mu}^{\mathcal{I}}, W^*)\big| \leq \big[L_r + H(1 + \lambda \log|\mathcal{A}|)L_P\big] \sum_{m=h}^H \int_0^1 \|\mu_m^\beta - \tilde{\mu}_m^\beta\|_1 \mathrm{d}\beta.$$

From the relationship between the value function and action-value function, we have that

$$\big|V_h^{\lambda,\alpha}(s, \pi^\alpha, \mu^{\mathcal{I}}, W^*) - V_h^{\lambda,\alpha}(s, \pi^\alpha, \tilde{\mu}^{\mathcal{I}}, W^*)\big| \leq \big[L_r + H(1 + \lambda \log|\mathcal{A}|)L_P\big] \sum_{m=h}^H \int_0^1 \|\mu_m^\beta - \tilde{\mu}_m^\beta\|_1 \mathrm{d}\beta.$$

Thus, we conclude the proof of Proposition 27. ■

Q.2.9 Proof of Proposition 17

Proof [Proof of Proposition 17] We first prove the claim related to the value function. From the definition of the optimal policy, we have that

$$V_h^{\lambda,\alpha}(s, \pi^{*,\mathcal{I}}, \mu^{\mathcal{I}}, W^*) = \max_{p \in \Delta(\mathcal{A})} \big\langle r_h\big(s, \cdot, z_h^\alpha(\mu_h^{\mathcal{I}}, W_h^*)\big), p \big\rangle - \lambda R(p)$$

$$+\sum_{a\in\mathcal{A}}\int_{\mathcal{S}}p(a)P_h\big(s'\,|\,s,a,z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)\big)V_{h+1}^{\lambda,\alpha}(s,\pi^{*,\mathcal{I}},\mu^{\mathcal{I}},W^*)\mathrm{d}s'.$$

Thus, we have that

$$\begin{aligned}
&\big|V_h^{\lambda,\alpha}(s,\pi^{*,\mathcal{I}},\mu^{\mathcal{I}},W^*)-V_h^{\lambda,\alpha}(s,\tilde{\pi}^{*,\mathcal{I}},\tilde{\mu}^{\mathcal{I}},W^*)\big|\\
&\quad\leq\big(H(1+\lambda\log|\mathcal{A}|)L_P+L_r\big)\int_0^1\|\mu_h^\beta-\tilde{\mu}_h^\beta\|_1\mathrm{d}\beta\\
&\qquad+\max_{s\in\mathcal{S}}\big|V_{h+1}^{\lambda,\alpha}(s,\pi^{*,\mathcal{I}},\mu^{\mathcal{I}},W^*)-V_{h+1}^{\lambda,\alpha}(s,\tilde{\pi}^{*,\mathcal{I}},\tilde{\mu}^{\mathcal{I}},W^*)\big|,
\end{aligned}$$

where the inequality results from the fact that $|\max_x f(x)-\max_x g(x)|\leq\max_x|f(x)-g(x)|$ and Assumption 1. By induction, it is easy to prove that

$$\begin{aligned}
&\max_{s\in\mathcal{S}}\big|V_h^{\lambda,\alpha}(s,\pi^{*,\mathcal{I}},\mu^{\mathcal{I}},W^*)-V_h^{\lambda,\alpha}(s,\tilde{\pi}^{*,\mathcal{I}},\tilde{\mu}^{\mathcal{I}},W^*)\big|\\
&\qquad\leq\big(H(1+\lambda\log|\mathcal{A}|)L_P+L_r\big)\sum_{m=h}^{H}\int_0^1\|\mu_m^\beta-\tilde{\mu}_m^\beta\|_1\mathrm{d}\beta.
\end{aligned}$$

Next, we prove the claim related to the optimal policies. From the definition of the optimal policies, we have that

$$\begin{aligned}
\pi_h^{*,\alpha}(\cdot\,|\,s)=\operatorname*{argmax}_{p\in\Delta(\mathcal{A})}&\big\langle r_h\big(s,\cdot,z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)\big),p\big\rangle-\lambda R(p)\\
&+\sum_{a\in\mathcal{A}}\int_{\mathcal{S}}p(a)P_h\big(s'\,|\,s,a,z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)\big)V_{h+1}^{\lambda,\alpha}(s,\pi^{*,\mathcal{I}},\mu^{\mathcal{I}},W^*)\mathrm{d}s'.
\end{aligned}$$

We define that

$$\begin{aligned}
y_h^\alpha(s,a)&=r_h\big(s,a,z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)\big)+\int_{\mathcal{S}}P_h\big(s'\,|\,s,a,z_h^\alpha(\mu_h^{\mathcal{I}},W_h^*)\big)V_{h+1}^{\lambda,\alpha}(s,\pi^{*,\mathcal{I}},\mu^{\mathcal{I}},W^*)\mathrm{d}s',\\
\tilde{y}_h^\alpha(s,a)&=r_h\big(s,a,z_h^\alpha(\tilde{\mu}_h^{\mathcal{I}},W_h^*)\big)+\int_{\mathcal{S}}P_h\big(s'\,|\,s,a,z_h^\alpha(\tilde{\mu}_h^{\mathcal{I}},W_h^*)\big)V_{h+1}^{\lambda,\alpha}(s,\pi^{*,\mathcal{I}},\tilde{\mu}^{\mathcal{I}},W^*)\mathrm{d}s'.
\end{aligned}$$

Then Lemma 34 shows that

$$\big\|\pi_h^{*,\alpha}(\cdot\,|\,s)-\tilde{\pi}_h^{*,\alpha}(\cdot\,|\,s)\big\|_1\leq\big\|y_h^\alpha(s,\cdot)-\tilde{y}_h^\alpha(s,\cdot)\big\|_\infty.$$

From the triangle inequality and Assumption 1, we have that

$$\begin{aligned}
&\big\|y_h^\alpha(s,\cdot)-\tilde{y}_h^\alpha(s,\cdot)\big\|_\infty\\
&\quad\leq\big(H(1+\lambda\log|\mathcal{A}|)L_P+L_r\big)\int_0^1\|\mu_h^\beta-\tilde{\mu}_h^\beta\|_1\mathrm{d}\beta\\
&\qquad+\big(H(1+\lambda\log|\mathcal{A}|)L_P+L_r\big)\sum_{m=h}^{H}\int_0^1\|\mu_m^\beta-\tilde{\mu}_m^\beta\|_1\mathrm{d}\beta,
\end{aligned}$$

which proves the claim related to the optimal policies. Thus, we conclude the proof of Proposition 17 ■

Q.3 Propositions and Lemmas for Combination

Q.3.1 Proof of Corollary 21

Proof [Proof of Corollary 21] Following the proof of Theorem 5, we decompose the risk difference as

$$\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h) - \mathcal{R}_{\bar{\xi}}(f_h^*, g_h^*, W_h^*) \\ = \text{Generalization Error of Risk} + \text{Empirical Risk Difference},$$

where the generalization error of risk and the empirical risk difference are defined similarly as those in Theorem 5. From the procedure of Algorithm (5), we have

$$\text{Empirical Risk Difference} \leq 0.$$

The generalization error of risk can be bounded by inequality (42) in the proof of Theorem 5. Thus, we conclude the proof of Corollary 21. ∎

Q.3.2 Proof of Proposition 22

.

Proof [Proof of Proposition 22] For any $h \in [H-1]$, the definition of Γ_2 shows that

$$\mu_{h+1}^{\alpha}(s') = \sum_{a\in\mathcal{A}} \int_{\mathcal{S}} P_h^*(s' \,|\, s, a, z_h^{\alpha}(\mu_h^{\mathcal{I}}, W_h^*))\mu_h^{\alpha}(s)\pi_h^{\alpha}(a \,|\, s)\mathrm{d}s.$$

Assumption 8 implies that we can bound the total variation between μ_{h+1}^{α} and $\hat{\mu}_{h+1}^{\alpha}$ as

$$\begin{aligned} &\|\mu_{h+1}^{\alpha} - \hat{\mu}_{h+1}^{\alpha}\|_1 \\ &\quad \leq L_{\varepsilon}\mathbb{E}_{\rho_h^{\alpha}}\Big[\big|\hat{f}_h\big(\omega_h^{\alpha}(\hat{W}_h)\big) - f_h^*\big(\omega_h^{\alpha}(W_h^*)\big)\big|\Big] + L_{\varepsilon}\mathbb{E}_{\rho_h^{\alpha}}\Big[\big|\hat{f}_h\big(\omega_h^{\alpha}(\hat{W}_h)\big) - \hat{f}_h\big(\tilde{\omega}_h^{\alpha}(\hat{W}_h)\big)\big|\Big] \\ &\qquad + \|\mu_h^{\alpha} - \hat{\mu}_h^{\alpha}\|_1, \end{aligned} \tag{94}$$

where

$$\tilde{\omega}_h^{\alpha}(W) = \int_0^1 \int_{\mathcal{S}} W(\alpha,\beta)k\big(\cdot, (s_{\tau,h}^i, a_{\tau,h}^i, s)\big)\hat{\mu}_h^{\beta}(s)\,\mathrm{d}s\,\mathrm{d}\beta.$$

The first term in the right-hand side of inequality (94) can be upper bounded as

$$L_{\varepsilon}\mathbb{E}_{\rho_h^{\alpha}}\Big[\big|\hat{f}_h\big(\omega_h^{\alpha}(\hat{W}_h) - f_h^*\big(\omega_h^{\alpha}(W_h^*)\big)\big|\Big] \leq e_h^{\pi,\alpha},$$

where the inequality results from the Hölder inequality. The second term in the right-hand side of inequality (94) can be upper-bounded as

$$L_{\varepsilon}\mathbb{E}_{\rho_h^{\alpha}}\Big[\big|\hat{f}_h\big(\omega_h^{\alpha}(\hat{W}_h)\big) - \hat{f}_h\big(\tilde{\omega}_h^{\alpha}(\hat{W}_h)\big)\big|\Big] \leq rL_K L_{\varepsilon}\|\omega_h^{\alpha}(\hat{W}_h) - \tilde{\omega}_h^{\alpha}(\hat{W}_h)\|_{\mathcal{H}} \leq rL_K L_{\varepsilon} B_k \int_0^1 \|\hat{\mu}_h^{\beta} - \mu_h^{\beta}\|_1 \mathrm{d}\beta,$$

where the first inequality results from Lemma 33, and the second inequality results from the triangle inequality. Thus, we have that

$$\|\mu_{h+1}^{\alpha} - \hat{\mu}_{h+1}^{\alpha}\|_1 \le \|\mu_h^{\alpha} - \hat{\mu}_h^{\alpha}\|_1 + rL_K L_\varepsilon B_k \int_0^1 \|\hat{\mu}_h^{\beta} - \mu_h^{\beta}\|_1 \mathrm{d}\beta + e_h^{\pi,\alpha}.$$

By induction, it is easy to prove that

$$\|\mu_h^{\alpha} - \hat{\mu}_h^{\alpha}\|_1 \le \sum_{m=1}^{h-2} \sum_{k=0}^{h-m-2} (1 + rL_K L_\varepsilon B_k)^k \int_0^1 e_m^{\pi,\beta} \mathrm{d}\beta + \sum_{m=1}^{h-1} e_m^{\pi,\alpha},$$

which proves our desired results. Thus, we conclude the proof of Proposition 22. ■

Q.3.3 Proof of Proposition 23

Proof [Proof of Proposition 23] From the definition of the action-value function, we have that

$$\begin{aligned} Q_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},W^*) &= r_h(s,a,z_h^\alpha(\mu^{\mathcal{I}},W_h^*)) + \mathbb{E}\big[V_{h+1}^{\lambda,\alpha}(s',\pi^\alpha,\mu^{\mathcal{I}},W^*) \,\big|\, s_h^\alpha = s, a_h^\alpha = a\big] \\ V_h^{\lambda,\alpha}(s',\pi^\alpha,\mu^{\mathcal{I}},W^*) &= \big\langle Q_h^{\lambda,\alpha}(s,\cdot,\pi^\alpha,\mu^{\mathcal{I}},W^*), \pi_h^\alpha(\cdot\,|\,s)\big\rangle + \lambda R\big(\pi_h^\alpha(\,\cdot\,|\,s)\big). \end{aligned} \tag{95}$$

Thus for any $h \in [H]$, we have that

$$\begin{aligned} &\mathbb{E}_{\rho_h^{+,\alpha}}\Big[\Big|\hat{Q}_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},\hat{W}) - Q_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},W^*)\Big|\Big] \\ &\quad \le \mathbb{E}_{\rho_h^{+,\alpha}}\Big[\Big|\hat{r}_h(s,a,z_h^\alpha(\mu^{\mathcal{I}},\hat{W}_h)) - r_h^*(s,a,z_h^\alpha(\mu^{\mathcal{I}},W_h^*))\Big|\Big] \\ &\qquad + (H-h)(1+\lambda\log|\mathcal{A}|)\mathbb{E}_{\rho_h^{+,\alpha}}\Big[\Big\|\hat{P}_h(\cdot\,|\,s,a,z_h^\alpha(\mu^{\mathcal{I}},\hat{W}_h)) - P_h^*(\cdot\,|\,s,a,z_h^\alpha(\mu^{\mathcal{I}},W_h^*))\Big\|_1\Big] \\ &\qquad + C\cdot\mathbb{E}_{\rho_{h+1}^{\mathrm{b},\alpha}}\Big[\Big|\hat{Q}_{h+1}^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},\hat{W}) - Q_{h+1}^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},W^*)\Big|\Big], \end{aligned}$$

where the inequality results from the triangle inequality and Eqn. (95). Since $\hat{Q}_{H+1}^{\lambda,\alpha} = Q_{H+1}^{\lambda,\alpha} = 0$, we have that

$$\begin{aligned} &\mathbb{E}_{\rho_h^{\mathrm{b},\alpha}}\Big[\Big|\hat{Q}_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},\hat{W}) - Q_h^{\lambda,\alpha}(s,a,\pi^\alpha,\mu^{\mathcal{I}},W^*)\Big|\Big] \\ &\quad \le \sum_{m=h}^{H} \mathbb{E}_{\rho_m^{\mathrm{b},\alpha}}\Big[\Big|\hat{g}_m\big(\omega_h^\alpha(\hat{W}_m)\big) - g_m^*\big(\omega_h^\alpha(W_m^*)\big)\Big|\Big] \\ &\qquad + L_\varepsilon H(1+\lambda\log|\mathcal{A}|)\sum_{m=h}^{H}\sqrt{\mathbb{E}_{\rho_m^{\mathrm{b},\alpha}}\Big[\Big|\hat{f}_m\big(\omega_h^\alpha(\hat{W}_m)\big) - f_m^*\big(\omega_h^\alpha(W_m^*)\big)\Big|\Big]}, \end{aligned}$$

where the inequality results from Assumption 8. Our desired result follows from the Hölder inequality. Thus, we conclude the proof of Proposition 23. ■

Q.3.4 Proof of Corollary 24

Proof [Proof of Corollary 24] We proof mainly takes two steps:

- Reformulate the algorithm (62).
- Decompose the risk difference and control each terms.

Step 1: Reformulate the algorithm (62)**.**
From the definition ϕ^*, we have that for $\alpha \in ((i-1)/N, i/N]$

$$\tilde{\mu}^{\alpha}_{\tau,h} = \mu^{\phi^*(i/N)+\alpha-i/N}_{\tau,h} = \mu^{\phi^*(\alpha)}_{\tau,h},$$

where the first equality results from the definition of $\tilde{\mu}^{\mathcal{I}}_{\tau}$, and the second equality results from that $\alpha \in ((i-1)/N, i/N]$ and ϕ^* is a permutation of $\{((i-1)/N, i/N]\}_{i=1}^N$. Thus, we have that

$$\begin{aligned}
\tilde{\omega}^i_{\tau,h}(W^{\phi}) &= \int_0^1 \int_{\mathcal{S}} W\big(\phi(i/N), \phi(\beta)\big) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big) \mu^{\phi^*(\beta)}_{\tau,h}(s)\, \mathrm{d}s\, \mathrm{d}\beta \\
&= \int_0^1 \int_{\mathcal{S}} W\big(\phi \circ \psi^* \circ \phi^*(i/N), \phi \circ \psi^* \circ \phi^*(\beta)\big) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big) \mu^{\phi^*(\beta)}_{\tau,h}(s)\, \mathrm{d}s\, \mathrm{d}\beta \\
&= \int_0^1 \int_{\mathcal{S}} W\big(\phi \circ \psi^*(\xi_i), \phi \circ \psi^*(\gamma)\big) k\big(\cdot, (s^i_{\tau,h}, a^i_{\tau,h}, s)\big) \mu^{\gamma}_{\tau,h}(s)\, \mathrm{d}s\, \mathrm{d}\gamma \\
&= \omega^i_{\tau,h}(W^{\phi\circ\psi^*}),
\end{aligned}$$

where the second equality results from that ϕ^* is the inverse functio of ψ^*, and the third equality results from taking $\gamma = \phi^*(\beta)$ and $\phi^*(i/N) = \xi_i$. Thus, Algorithm (62) can be equivalent formulated as

$$\begin{aligned}
&(\hat{f}_h, \hat{g}_h, \hat{W}_h, \hat{\phi}_h) \\
&\quad = \mathop{\mathrm{argmin}}_{\substack{f\in\mathbb{B}(r,\tilde{\mathcal{H}}) \\ g\in\mathbb{B}(\tilde{r},\tilde{\mathcal{H}}) \\ W\in\tilde{\mathcal{W}} \\ \phi\in\mathcal{C}^N_{[0,1]}}} \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \Big(s^i_{\tau,h+1} - f\big(\omega^i_{\tau,h}(W^{\phi\circ\psi^*})\big)\Big)^2 + \Big(r^i_{\tau,h} - g\big(\omega^i_{\tau,h}(W^{\phi\circ\psi^*})\big)\Big)^2.
\end{aligned} \tag{96}$$

Step 2: Decompose the risk difference and control each terms.

$$\mathcal{R}_{\bar{\xi}}(\hat{f}_h, \hat{g}_h, \hat{W}_h^{\hat{\phi}_h\circ\psi^*}) - \mathcal{R}_{\bar{\xi}}(f^*_h, g^*_h, W^*_h) = \text{Generalization Error of Risk} + \text{Empirical Risk Difference},$$

where the generalization error of risk and the empirical risk difference are defined as

$$\begin{aligned}
&\text{Generalization Error of Risk} \\
&\quad = \frac{1}{NL} \sum_{\tau=1}^{L} \sum_{i=1}^{N} \mathbb{E}_{\rho^i_{\tau,h}} \bigg[\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\bigg]
\end{aligned}$$

$$- 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2$$
$$+ \frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\mathbb{E}_{\rho^i_{\tau,h}}\Big[\Big(r^i_{\tau,h} - \hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(r^i_{\tau,h} - g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2\Big]$$
$$- 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(r^i_{\tau,h} - \hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(r^i_{\tau,h} - g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2,$$

Empirical Risk Difference

$$= 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(s^i_{\tau,h+1} - \hat{f}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(s^i_{\tau,h+1} - f^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2$$
$$+ 2\frac{1}{NL}\sum_{\tau=1}^{L}\sum_{i=1}^{N}\Big(r^i_{\tau,h} - \hat{g}_h\big(\omega^i_{\tau,h}(\hat{W}_h^{\hat{\phi}_h\circ\psi^*})\big)\Big)^2 - \Big(r^i_{\tau,h} - g^*_h\big(\omega^i_{\tau,h}(W^*_h)\big)\Big)^2.$$

From the procedure of Algorithm (96), we have

$$\text{Empirical Risk Difference} \leq 0.$$

The generalization error of risk can be controlled exactly as the proof of Theorem 7. Thus, we conclude the proof of Corollary 24. ■

Q.4 Technical Lemmas

Lemma 31 *For a finite alphabet $\mathcal{X}$ and any distribution p supported on it, we define $p_\beta = (1-\beta)p + \beta\,\mathit{Unif}(\mathcal{X})$. Then the function $f(\beta) = R(p_\beta)$ is a decreasing function on $\beta \in [0,1]$.*

Proof [Proof of Lemma 31] From calculus, we can show that

$$f''(\beta) \geq 0 \text{ for } \beta \in [0,1].$$

Since $f'(1) = 0$, we have that $f'(\beta) \leq 0$ for $\beta \in [0,1]$. Thus, we conclude the proof of Lemma 31. ■

Lemma 32 (Theorem 3.5 in Pinelis (1994)) *Let $X_1, \cdots, X_n$ be independent random variables that take values in a Hilbert space. If $\|X_i\| \leq M$ and $\mathbb{E}[X_i] = 0$ for all $i \in [n]$. Then $\mathbb{P}(\|X_1 + \cdots + X_n\| \geq t) \leq 2\exp(-t^2/(2nM))$.*

Lemma 33 *In a RKHS $\mathcal{H}$ with kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ that satisfies: (i) $k(x,x) \leq B_k^2$ for all $x \in \mathcal{X}$. (ii) $\|k(\cdot,x) - k(\cdot,x')\|_{\mathcal{H}} \leq L_k\|x - x'\|_{\mathcal{X}}$ for all $x, x' \in \mathcal{X}$. We have that for any $f \in \mathbb{B}(r,\mathcal{H})$: (i) $|f(x)| \leq rB_k$ for all $x \in \mathcal{X}$. (ii) $|f(x) - f(x')| \leq rL_k\|x - x'\|_{\mathcal{X}}$ for all $x, x' \in \mathcal{X}$.*

Proof [Proof of Lemma 33] For the first claim, we have that

$$\big|f(x)\big| = \Big|\big\langle f, k(\cdot, x)\big\rangle\Big| \le \|f\|_{\mathcal{H}} \cdot \big\|k(\cdot, x)\big\|_{\mathcal{H}} \le rB_k.$$

For the second claim, we have that

$$\big|f(x) - f(x')\big| = \Big|\big\langle f, k(\cdot, x) - k(\cdot, x')\big\rangle\Big| \le \|f\|_{\mathcal{H}} \cdot \big\|k(\cdot, x) - k(\cdot, x')\big\|_{\mathcal{H}} \le rL_k\|x - x'\|_{\mathcal{X}}.$$

Thus, we conclude the proof of Lemma 33. ■

Lemma 34 *Let $\mathcal{X}$ be a nonempty compact convex set and $f : \mathcal{X} \to \mathbb{R}$ be a differentiable k-strongly convex function, i.e., $f(x) \ge f(y) + \langle \nabla f(y), x - y\rangle + \frac{k}{2}\|x - y\|^2$ for all $x, y \in \mathcal{X}$. For any two elements y_1, y_2, we define*

$$x_i = \operatorname*{argmax}_{x\in\mathcal{X}} \langle x, y_i\rangle - f(x) \text{ for } i = 1, 2.$$

Then $\|x_1 - x_2\| \le \|y_1 - y_2\|_*/k$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Proof [Proof of Lemma 34] Define $f_i(x) = \langle x, y_i\rangle - f(x)$ for $i = 1, 2$. Then Shalev-Shwartz (2012, Lemma 2.8) shows that

$$\frac{k}{2}\|x_1 - x_2\|^2 \le f_1(x_1) - f_1(x_2) \quad \text{and} \quad \frac{k}{2}\|x_1 - x_2\|^2 \le f_2(x_2) - f_2(x_1).$$

Summing these two inequalities, we have

$$k\|x_1 - x_2\|^2 \le \langle x_1 - x_2, y_1 - y_2\rangle \le \|x_1 - x_2\| \cdot \|y_1 - y_2\|_*,$$

where the second inequality results from the definition of the dual norm. Thus, we conclude the proof of Lemma 34. ■

Lemma 35 (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*

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

Lemma 36 *For any two distributions $p^*, p \in \Delta(\mathcal{A})$ and $\hat{p} = (1 - \beta)p + \beta\,\mathrm{Unif}(\mathcal{A})$ with $\beta \in (0, 1)$. Then*

$$\mathrm{KL}(p^*\|\hat{p}) \le \log\frac{|\mathcal{A}|}{\beta}$$

$$\mathrm{KL}(p^*\|\hat{p}) - \mathrm{KL}(p^*\|p) \le \beta/(1 - \beta).$$

Proof [Proof of Lemma 36]

$$\mathrm{KL}(p^*\|\hat{p}) \le \Big\langle p^*, \log\frac{p^*}{(1-\beta)p+\beta/|\mathcal{A}|}\Big\rangle \le \Big\langle p^*, \log\frac{1}{\beta/|\mathcal{A}|}\Big\rangle = \log\frac{|\mathcal{A}|}{\beta}.$$

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

$$\mathrm{KL}(p^*\|\hat{p}) - \mathrm{KL}(p^*\|p) = \Big\langle p^*, \log\frac{p}{(1-\beta)p+\beta/|\mathcal{A}|}\Big\rangle \le \Big\langle p^*, \log\frac{p}{(1-\beta)p}\Big\rangle \le \Big\langle p^*, \frac{\beta}{1-\beta}\Big\rangle = \frac{\beta}{1-\beta},$$

where the second inequality results from $\log(x) \le x-1$ for $x > 0$. Thus, we conclude the proof of Lemma 36. ∎

Lemma 37 (Performance Difference Lemma) *Given a policy $\pi^{\mathcal{I}}$ and the corresponding mean-field flow $\mu^{\mathcal{I}}$, for any agent $\alpha \in \mathcal{I}$ and any policy $\tilde{\pi}^\alpha$, we have*

$$\begin{aligned} &V_1^{\lambda,\alpha}(s,\tilde{\pi}^\alpha,\mu^{\mathcal{I}},W) - V_1^{\lambda,\alpha}(s,\pi^\alpha,\mu^{\mathcal{I}},W) + \lambda\mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\bigg[\sum_{h=1}^H \mathrm{KL}\big(\tilde{\pi}_h^\alpha(\cdot\,|\,s_h^\alpha)\|\pi_h^\alpha(\cdot\,|\,s_h^\alpha)\big)\,\Big|\, s_1^\alpha = s\bigg] \\ &\quad = \mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\bigg[\sum_{h=1}^H \big\langle Q_h^{\lambda,\alpha}(s_h^\alpha,\cdot,\pi^\alpha,\mu^{\mathcal{I}},W) - \lambda\log\pi_h^\alpha(\cdot\,|\,s_h^\alpha), \tilde{\pi}_h^\alpha(\cdot\,|\,s_h^\alpha) - \pi_h^\alpha(\cdot\,|\,s_h^\alpha)\big\rangle \,\Big|\, s_1^\alpha = s\bigg], \end{aligned}$$

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 agent α under the MDP induced by $\mu^{\mathcal{I}}$.

Proof [Proof of Lemma 37] From the definition of $V_1^{\lambda,\alpha}(s,\tilde{\pi}^\alpha,\mu^{\mathcal{I}},W)$, we have

$$\begin{aligned} &V_1^{\lambda,\alpha}(s,\tilde{\pi}^\alpha,\mu^{\mathcal{I}},W) \\ &\quad = \mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\bigg[\sum_{h=1}^H r_h(s_h^\alpha,a_h^\alpha,z_h^\alpha) - \lambda\log\tilde{\pi}_h^\alpha(a_h^\alpha|s_h^\alpha) + V_h^{\lambda,\alpha}(s_h^\alpha,\pi^\alpha,\mu^{\mathcal{I}},W) - V_h^{\lambda,\alpha}(s_h^\alpha,\pi^\alpha,\mu^{\mathcal{I}},W)\bigg| s_1^\alpha = s\bigg] \\ &\quad = \mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\bigg[\sum_{h=1}^H r_h(s_h^\alpha,a_h^\alpha,z_h^\alpha) - \lambda\log\tilde{\pi}_h^\alpha(a_h^\alpha\,|\,s_h^\alpha) + V_{h+1}^{\lambda,\alpha}(s_{h+1}^\alpha,\pi^\alpha,\mu^{\mathcal{I}},W) \\ &\qquad - V_h^{\lambda,\alpha}(s_h^\alpha,\pi^\alpha,\mu^{\mathcal{I}},W)\,\bigg|\, s_1^\alpha = s\bigg] + V_1^{\lambda,\alpha}(s,\pi^\alpha,\mu^{\mathcal{I}},W), \end{aligned} \tag{97}$$

where the second equality results from the rearrangement from the terms. We then focus on a part of the right-hand side of Eqn. (97).

$$\begin{aligned} &\mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\big[r_h(s_h^\alpha,a_h^\alpha,z_h^\alpha) - \lambda\log\tilde{\pi}_h^\alpha(a_h^\alpha\,|\,s_h^\alpha) + V_{h+1}^{\lambda,\alpha}(s_{h+1}^\alpha,\pi^\alpha,\mu^{\mathcal{I}},W)\,|\,s_1^\alpha = s\big] \\ &\quad = \mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\big[r_h(s_h^\alpha,a_h^\alpha,z_h^\alpha) + V_{h+1}^{\lambda,\alpha}(s_{h+1}^\alpha,\pi^\alpha,\mu^{\mathcal{I}},W)\,|\,s_1^\alpha = s\big] - \lambda\mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\Big[R\big(\tilde{\pi}_h^\alpha(\cdot\,|\,s_h^\alpha)\big)\,|\,s_1^\alpha = s\Big] \\ &\quad = \mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\Big[\big\langle Q_h^{\lambda,\alpha}(s_h^\alpha,\cdot,\pi^\alpha,\mu^{\mathcal{I}},W), \tilde{\pi}_h^\alpha(\cdot\,|\,s_h^\alpha)\big\rangle\,|\,s_1^\alpha = s\Big] - \lambda\mathbb{E}_{\tilde{\pi}^\alpha,\mu^{\mathcal{I}}}\Big[R\big(\tilde{\pi}_h^\alpha(\cdot\,|\,s_h^\alpha)\big)\,|\,s_1^\alpha = s\Big], \end{aligned} \tag{98}$$

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 equality results from the definition of $Q_h^{\lambda,\alpha}$ and $V_{h+1}^{\lambda,\alpha}$. Substituting Eqn. (98) into Eqn. (97) and noting the fact that $V_h^{\lambda,\alpha}(s_h^\alpha, \pi^\alpha, \mu^{\mathcal{I}}, W) = \langle Q_h^{\lambda,\alpha}(s_h^\alpha, \cdot, \pi^\alpha, \mu^{\mathcal{I}}, W), \pi_h^\alpha(\cdot \,|\, s_h^\alpha)\rangle - R(\pi_h^\alpha(\cdot \,|\, s_h^\alpha))$, we derive that

$$
\begin{aligned}
&V_1^{\lambda,\alpha}(s, \tilde{\pi}^\alpha, \mu^{\mathcal{I}}, W) - V_1^{\lambda,\alpha}(s, \pi^\alpha, \mu^{\mathcal{I}}, W) \\
&\quad = \mathbb{E}_{\tilde{\pi}^\alpha, \mu^{\mathcal{I}}}\Bigg[\sum_{h=1}^H \big\langle Q_h^{\lambda,\alpha}(s_h^\alpha, \cdot, \pi^\alpha, \mu^{\mathcal{I}}, W), \tilde{\pi}_h^\alpha(\cdot \,|\, s_h^\alpha) - \pi_h^\alpha(\cdot \,|\, s_h^\alpha)\big\rangle \,|\, s_1^\alpha = s\Bigg] \\
&\qquad - \lambda \mathbb{E}_{\tilde{\pi}^\alpha, \mu^{\mathcal{I}}}\Bigg[\sum_{h=1}^H R\big(\tilde{\pi}_h^\alpha(\cdot \,|\, s_h^\alpha)\big) - R\big(\pi_h^\alpha(\cdot \,|\, s_h^\alpha)\big) \,|\, s_1^\alpha = s\Bigg] \\
&\quad = \mathbb{E}_{\tilde{\pi}^\alpha, \mu^{\mathcal{I}}}\Bigg[\sum_{h=1}^H \big\langle Q_h^{\lambda,\alpha}(s_h^\alpha, \cdot, \pi^\alpha, \mu^{\mathcal{I}}, W), \tilde{\pi}_h^\alpha(\cdot \,|\, s_h^\alpha) - \pi_h^\alpha(\cdot \,|\, s_h^\alpha)\big\rangle \,|\, s_1^\alpha = s\Bigg] \\
&\qquad - \lambda \mathbb{E}_{\tilde{\pi}^\alpha, \mu^{\mathcal{I}}}\Bigg[\sum_{h=1}^H \mathrm{KL}\big(\tilde{\pi}_h^\alpha(\cdot|s_h^\alpha) \| \pi_h^\alpha(\cdot|s_h^\alpha)\big) + \big\langle \log \pi_h^\alpha(\cdot|s_h^\alpha), \tilde{\pi}_h^\alpha(\cdot|s_h^\alpha) - \pi_h^\alpha(\cdot|s_h^\alpha)\big\rangle \,|\, s_1^\alpha = s\Bigg],
\end{aligned}
$$

where the last equality results from the definition of the negative entropy $R(\cdot)$. This concludes the proof of Lemma 37.

■

Lemma 38 *For a finite alphabet $\mathcal{X}$, define R as the negative entropy function. For two distributions p, q supported on $\mathcal{X}$, we have that*

$$
|R(p) - R(q)| \le \max\Big\{\big\|\log(p)\big\|_\infty, \big\|\log(q)\big\|_\infty\Big\}\|p - q\|_1.
$$

Proof [Proof of Lemma 38] Then we have that

$$
|R(p) - R(q)| \le \int_0^1 \Big|\Big\langle \nabla R\big(q + t(p-q)\big), p - q\Big\rangle\Big| \mathrm{d}t \le \|p - q\|_1 \int_0^1 \Big\|\log\big(q + t(p-q)\big)\Big\|_\infty \mathrm{d}t,
$$

where the first inequality results from the definition of integral and the triangle inequality, and the second inequality results from Hölder's inequality. The desired result follows from the fact that for $t \in [0, 1]$

$$
\Big\|\log\big(q + t(p-q)\big)\Big\|_\infty \le \max\Big\{\big\|\log(p)\big\|_\infty, \big\|\log(q)\big\|_\infty\Big\}.
$$

Thus, we conclude the proof of Lemma 38. ■

Lemma 39 (Lemma 3 in Xie et al. (2021)) *Let $p, q, u \in \Delta(\mathcal{X})$ be distributions supported on a finite set $\mathcal{X}$. If $p(x) \ge \alpha_1$, $q(x) \ge \alpha_1$, and $u(x) \ge \alpha_2$ for all $x \in \mathcal{X}$. Then*

$$
\mathrm{KL}(p\|u) - \mathrm{KL}(q\|u) \le \bigg(1 + \log \frac{1}{\min\{\alpha_1, \alpha_2\}}\bigg)\|p - q\|_1
$$

Lemma 40 (Lemma 39 in Wei et al. (2021)) *Let* $\{g_t\}_{t\geq 0}$ *and* $\{h_t\}_{t\geq 0}$ *be non-negative sequences that satisft* $g_t \leq (1-c)g_{t-1} + h_t$ *for some* $0 < c < 1$ *for all* $t \geq 1$*. Then*

$$g_t \leq g_0(1-c)^t + \frac{\max_{\tau\in[1,t/2]} h_\tau}{c}(1-c)^{t/2} + \frac{\max_{\tau\in[t/2,t]} h_\tau}{c}.$$

References

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR, 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.

J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

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.

P. E. Caines and M. Huang. Graphon mean field games and their equations. *SIAM Journal on Control and Optimization*, 59(6):4373–4399, 2021.

P. Cardaliaguet and S. Hadikhanloo. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

R. Carmona, M. Laurière, and Z. Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *arXiv preprint arXiv:1910.12802*, 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.

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.

K. Cui and H. Koeppl. Learning graphon mean field games and approximate Nash equilibria. *International Conference on Learning Representations*, 2021b.

K. Elamvazhuthi and S. Berman. Mean-field models in swarm robotics: A survey. *Bioinspiration & Biomimetics*, 15(1):015001, 2019.

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

Z. Fang, Z. Guo, and D. Zhou. Optimal learning rates for distribution regression. *Journal of Complexity*, 56:101426, 2020.

C. Gao and Z. Ma. Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statistical Science*, 36:16–33, 2021.

C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43:2624–2652, 2015.

S. Gao and P. E. Caines. Graphon control of large-scale networks of linear systems. *IEEE Transactions on Automatic Control*, 65(10):4090–4105, 2019.

S. Gao, R. F. Tchuendom, and P. E. Caines. Linear quadratic graphon field games. *arXiv preprint arXiv:2006.03964*, 2020.

S. Gao, P. E. Caines, and M. Huang. Lqg graphon mean field games: Graphon invariant subspaces. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 5253–5260. IEEE, 2021.

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.

S. Gronauer and K. Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, pages 1–49, 2022.

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

X. Guo, A. Hu, and J. Zhang. MF-OMO: An optimization formulation of mean-field games. *arXiv preprint arXiv:2206.09608*, 2022a.

X. Guo, R. Xu, and T. Zariphopoulou. Entropy regularization for mean field games with learning. *Mathematics of Operations research*, 47(4):3239–3260, 2022b.

X. Guo, A. Hu, R. Xu, and J. Zhang. A general framework for learning mean-field games. *Mathematics of Operations Research*, 48(2):656–686, 2023.

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

S. Hadikhanloo. Learning in anonymous nonatomic games with applications to first-order mean field games. *arXiv preprint arXiv:1704.00378*, 2017.

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.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

C. Jin, Q. Liu, and S. Miryoosefi. Bellman Eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418, 2021.

N. Kallus, Y. Saito, and M. Uehara. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pages 5247–5256. PMLR, 2021.

O. Klopp and N. Verzelen. Optimal graphon estimation in cut distance. *Probability Theory and Related Fields*, 174:1033–1090, 2019.

O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45:316–354, 2017.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, pages 1–48, 2022.

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

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

P. Lavigne and L. Pfeiffer. Generalized conditional gradient and learning in potential mean field games. *arXiv preprint arXiv:2209.12772*, 2022.

K. Menda, Y. Chen, J. Grana, J. W. Bono, B. D. Tracey, M. J. Kochenderfer, and D. Wolpert. Deep reinforcement learning for event-driven multi-agent decision processes. *IEEE Transactions on Intelligent Transportation Systems*, 20(4):1259–1268, 2018.

D. Meunier, M. Pontil, and C. Ciliberto. Distribution regression with sliced wasserstein kernels. In *International Conference on Machine Learning*, pages 15501–15523. PMLR, 2022.

O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Oroojlooy and D. Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, pages 1–46, 2022.

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

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

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.

I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

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.

Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pages 948–957. PMLR, 2015.

Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Y. Tang and D. Ha. The sensory neuron as a transformer: Permutation-invariant neural networks for reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

R. F. Tchuendom, P. E. Caines, and M. Huang. On the master equation for linear quadratic graphon mean field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1026–1031. IEEE, 2020.

M. Uehara, J. Huang, and N. Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

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.

D. Wang, R. Walters, and R. Platt. SO(2)-equivariant reinforcement learning. *arXiv preprint arXiv:2203.04439*, 2022a.

D. Wang, R. Walters, X. Zhu, and R. Platt. Equivariant Q-Learning in Spatial Action Spaces. In *Conference on Robot Learning*, pages 1713–1723. PMLR, 2022b.

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.

C. Wei, C. Lee, M. Zhang, and H. Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on Learning Theory*, pages 4259–4299. PMLR, 2021.

P. J. Wolfe and S. C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

X. Xia, G. Mishne, and Y. Wang. Implicit graphon neural representation. In *International Conference on Artificial Intelligence and Statistics*, pages 10619–10634. PMLR, 2023.

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, 2021.

J. Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pages 5433–5442. PMLR, 2018.

K. Xu, Y. Zhang, D. Ye, P. Zhao, and M. Tan. Relation-aware transformer for portfolio policy learning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4647–4653, 2021.

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

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.

W. Zhan, B. Huang, A. Huang, N. Jiang, and J. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

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