Welfare Maximization in Competitive Equilibrium: Reinforcement Learning for Markov Exchange Economy

Zhihan Liu [*1] **Miao Lu** [*2] **Zhaoran Wang** [1] **Michael I. Jordan** [3] **Zhuoran Yang** [4]

Abstract

We study a bilevel economic system, which we refer to as a *Markov exchange economy* (MEE), from the point of view of multi-agent reinforcement learning (MARL). An MEE involves a central planner and a group of self-interested agents. The goal of the agents is to form a Competitive Equilibrium (CE), where each agent myopically maximizes her own utility at each step. The goal of the central planner is to steer the system so as to maximize social welfare, which is defined as the sum of the utilities of all agents. Working in a setting in which the utility function and the system dynamics are both unknown, we propose to find the socially optimal policy and the CE from data via both online and offline variants of MARL. Concretely, we first devise a novel suboptimality metric specifically tailored to MEE, such that minimizing such a metric certifies globally optimal policies for both the planner and the agents. Second, in the online setting, we propose an algorithm, dubbed as `MOLM`, which combines the optimism principle for exploration with subgame CE seeking. Our algorithm can readily incorporate general function approximation tools for handling large state spaces and achieves a sublinear regret. Finally, we adapt the algorithm to an offline setting based on the pessimism principle and establish an upper bound on the suboptimality.

[*]Equal contribution [1]Department of Industrial Engineering and Management Sciences, Northwestern University [2]School of the Gifted Young, University of Science and Technology of China [3]Department of Statistics, University of California, Berkeley [4]Department of Statistics and Data Science, Yale University. Correspondence to: Zhihan Liu <zhihanliu2027@u.northwestern.edu>, Miao Lu <lumiao@mail.ustc.edu.cn>, Zhaoran Wang <zhaoranwang@gmail.com>, Zhuoran Yang <zhuoran.yang@yale.edu>, Michael I. Jordan <jordan@cs.berkeley.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Many real-world economic systems involve interactions between a central planner and a group of self-interested agents, where the planner aims to find a policy that steers the agents to some ideal equilibrium that maximizes social welfare. One widely studied instance is optimal tax policy design (Mirrlees, 1976; Mankiw et al., 2009), where the tax policy-maker aims at balancing equality and productivity for tax-payers in the society. Less studied in the previous literature, the design of learning mechanisms for a bilevel economic system remains challenging due to the instability and co-adaptation between agents and the planner, especially for sequential decision-making problems. Despite the progress shown by several works (Kutschinski et al., 2003; Mannion et al., 2016; Zheng et al., 2020; 2021; Lussange et al., 2021) that apply multi-agent reinforcement learning (MARL) to instances of economic systems, it is still an open theoretical challenge to design efficient mechanisms for bilevel economic systems with provable guarantees.

Our approach brings MARL methods together with the classic model exchange economy (EE). The EE framework has a wide range of applications, including ride-sharing, operations management, crowdsourcing, wireless networks, and compute clusters (Cohen & Cyert, 1965; Hussain et al., 2013; Dissanayake et al., 2015; Rauch & Schleicher, 2015). In an exchange economy, a set of rational agents with individual initial endowments allocate and exchange a finite set of valuable resources based on a common price system. The target of EE is to achieve Competitive Equilibrium (CE), where all agents maximize their own utility under their budget constraint. Adapted from EE, our proposed framework, the *Markov exchange economy* (MEE), comprises a central planner, multiple agents, and contextual states which follow a Markov Decision Process (MDP). In MEE, the agents follow the same procedure as in EE conditioned on a contextual state. The central planner's action affects the evolution of the endowments of the agents as well as the contextual states. The goal of each agent is to myopically maximize its own utility at each step, which leads to a Competitive Equilibrium (CE) as the agents' subproblem. The goal of the central planner is to steer the system so as to achieve social welfare maximization (SWM), where social welfare

is defined as the sum of the utilities of all agents over the entire episode. Instead of adding restrictive assumptions that utility functions are known as in many prior works on EE (Tiwari et al., 2009; Hindman et al., 2011; Dissanayake et al., 2015), we aim to solve MEEs when learning both the unknown utility functions and transitions. In reality, it is difficult to collect an exact utility function through automated systems (Hindman et al., 2011; Delimitrou & Kozyrakis, 2013; Venkataraman et al., 2016; Rzadca et al., 2020; Guo et al., 2021) and assuming the full knowledge of transition probability is also unrealistic, which makes the problem still more challenging.

Taking one specific example for illustration, we consider a developer community. In this community, there are multiple developers who wish to myopically maximize their utility and an administrator who plans to maximize the sum of these developers' utilities. Each developer has her own endowments, e.g., computing resources, memory, bandwidth, and programmer time, for exchange within the community that are available for a finite number of timesteps. At each timestep, developers report their utilities based on their current allocations and contextual states (electricity fee or available time for device usage) to the administrator through rating systems. Meanwhile, the administrator implements a regulatory regime based on the collected utilities and current contextual state. The transition probability of the next contextual state is only determined by administrator's conducted regulation and current contextual state.

In this paper, we advocate MARL as a principled method for solving MEE. When interaction with environment is accessible, we learn the policies of the agents and the planner through online MARL methods. When only a historical dataset is available, we turn to an offline MARL protocol. To this end, we focus on the following question.

Can we design provably efficient online and offline algorithms for learning the policies of the planner and agents to achieve CE and SWM simultaneously in MEE?

Several challenges arise when addressing this question. First, from a theoretical point of view, it remains unknown how to mathematically characterize the jointly optimal policy of a planner and agents such that we can directly measure the performance of any planner-agent policy in terms of SWM while achieving CE among agents. Secondly, in the online and offline settings, where the MEE model is not known a priori, it remains unknown how to find the optimal policy for both planner and agents when this is coupled with the problem of balancing the exploration-exploitation tradeoff in an online setting and the problem of distribution shift in an offline setting. Finally, there are generally infinitely many states since the endowments of agents can be continuous, and it is unknown how to handle large state spaces in such online and offline learning problems, especially when the utilities and transitions are of general functional forms.

Our work addresses these challenges and provides an affirmative answer to the desired question. Specifically, by characterizing the optimal policy of planner and agents via a fixed-point formulation, we devise a novel suboptimality metric such that the suboptimality being zero is equivalent to the planner-agent policy being jointly optimal. Then, for the online setting where we learn the optimal policy by interacting with the MEE, we propose a model-based MARL algorithm, dubbed as `MOLM`, which combines the Optimism in Face of Uncertainty (OFU) principle (Auer et al., 2002; 2009; Jin et al., 2018; 2019) with a subroutine which solves the subgame CE for the agents at each timestep. Our algorithm can readily incorporate general function approximators such as kernel functions and neural networks in the estimation of the transition model and is shown to achieve a sublinear regret with respect to the newly designed suboptimality metric. Furthermore, for the offline setting where we aim to learn the optimal policy solely from a given dataset, we propose a similar algorithm that incorporates the pessimism principle (Buckman et al., 2020; Jin et al., 2021b) to overcome the distributional shift between trajectories in the dataset and those induced by the optimal policy. This algorithm is also able to employ general function approximators and is shown to find a policy whose suboptimality decays sublinearly in the size of the dataset. Finally, as a byproduct, we prove that our algorithms achieve approximately fair division among the agents (Varian, 1973; Budish et al., 2017; Babaioff et al., 2019) in both the online and offline settings.

Contributions. Our contributions are three-fold. First, we propose a new economic system known as MEE in attempt to understand the theoretical properties of solutions to planner-agent economic systems via MARL approaches. We define a suboptimality function to characterize the optimal policy for the planner and the agents in an MEE, with another suboptimality proposed to characterize the fair division property among the agents. Second, we design a MARL-style algorithm `MOLM` to find the optimal policy for the planner and the agents from data in online setting. For `MOLM` we establish an online regret upper bound, $\widetilde{\mathcal{O}}(\sqrt{dH^4N^2K})$, where K is the number of episodes, H is the time step, N is the number of agents, d is the eluder dimension of the general function class used by `MOLM`, and $\widetilde{\mathcal{O}}(\cdot)$ hides the logarithmic terms and constants. Third, in addition to `MOLM`, we design `MPLM` for offline MEE. For `MPLM`, we establish an offline suboptimality bound, $\widetilde{\mathcal{O}}(\sqrt{C^\star_\rho H^4N^2/K})$, where K is the size of dataset and $C^\star_\rho$ is the distribution shift coefficient in sense of partial coverage. Theoretical results show that both `MOLM` and `MPLM` provably find the optimal policy for planner and agents in the two settings. In addition, they provably achieve

fair division among agents as a byproduct.

1.1. Related Work

Our work adds to the line of research in applying machine learning methods to economic problems such as EE (Guo et al., 2021), mechanism design (Kandasamy et al., 2020), and social planning problems (Blaug, 2007). Our analysis of MEE is based on previous works on EE (Debreu, 1982; Zhang, 2011). Motivated by the extensive literature on online and offline RL, our works apply the optimism principle (Auer et al., 2002; Jin et al., 2018) and pessimism principle (Buckman et al., 2020; Jin et al., 2021b) in online and offline settings, respectively. Our work is also related to literature in MARL (Bucarey et al., 2019; Zhong et al., 2021) and RL with general function approximations (Xie et al., 2021; Cai et al., 2020b). However, none of the previous work analyzes bilevel economic systems, as we do for MEE in this paper. See Appendix B for full discussions of related work.

Notations We provide a table of notation in Appendix A.

2. Preliminaries

In this section, we introduce our economic model known as **M**arkovian **E**xchange **E**conomy (MEE) which involves several self-interested agents and a social planner. We specify the goal for both planner and agents, and we characterize their jointly optimal policy via a fixed-point formulation. All the proofs for the theorems are referred to Appendix E.

2.1. Markovian Exchange Economy

We define a finite horizon Markovian exchange economy as $(\mathcal{S}, \mathcal{A}, \mathcal{B}, N, L, H, \{u_h^{(i)}\}_{i\in[N],h\in[H]}, \{P_h\}_{h\in[H]})$ which consists of N agents, one social planner, L goods, and H time steps. The state space is denoted by $\mathcal{S} = \mathcal{C} \times \mathcal{E}^N$, where $\mathcal{C}$ is the context space and $\mathcal{E} \subseteq [0,1]^L$ is the space of each agent's endowments. A state at step h is denoted by $s_h = (c_h, e_h^{(1)}, \cdots, e_h^{(N)}) \in \mathcal{S}$. The agents' action space is denoted by $\mathcal{A} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(N)} \times [0,1]^L$, where $\mathcal{X}^{(i)} \in [0,1]^L$ is the allocation space of the i^{th} agent and $[0,1]^L$ is the price space. We denote by $a_h = (x_h^{(1)}, \cdots, x_h^{(N)}, p_h)$ the agents' action at step h. The planner's action space is denoted by $\mathcal{B}$ which is discrete, and the planner's action at step h is denoted by $b_h \in \mathcal{B}$. The utility function of the i^{th} agent at step h is denoted by $u_h^{(i)} : \mathcal{S} \times \mathcal{X}^{(i)} \mapsto [0,1]$. The transition kernel at step h is denoted by $P_h(s'|s,b) : \mathcal{S} \times \mathcal{B} \mapsto \Delta(\mathcal{S})$. We note that the transition kernel P_h does not depend on the agents' action, but only the planner's.

Policy and Value Functions. Without loss of generality, in the sequel we always focus on deterministic policies for both planner and agents. A planner's policy is denoted by $\pi = \{\pi_h\}_{h\in[H]}$ where $\pi_h : \mathcal{S} \mapsto \mathcal{B}$. An agents' policy is denoted by $\nu = \{\nu_h\}_{h\in[H]}$ where $\nu_h : \mathcal{S} \mapsto \mathcal{A}$, $s \mapsto (\nu_h^{(1)}(s), \cdots, \nu_h^{(N)}(s), \nu_h^{\mathbf{P}}(s))$. That is, $\nu_h^{(i)}$ determines the allocation of the i^{th} agent and $\nu_h^{\mathbf{P}}$ determines the price. We assume that π and ν belong to classes $\mathbf{\Pi}$ and $\mathbf{N}$ respectively. Given any pair of policy (π, ν), we define its action-value function and state value function recursively as

$$Q_h^{(\pi,\nu),(i)}(s_h, x_h^{(i)}, b_h) = u_h^{(i)}(s_h, x_h^{(i)}) + \int_{\mathcal{S}} V_{h+1}^{(\pi,\nu),(i)}(s') P_h(\mathrm{d}s'|s_h, b_h), \quad (1)$$
$$V_h^{(\pi,\nu),(i)}(s_h) = Q_h^{(\pi,\nu),(i)}(s_h, \nu_h^{(i)}(s_h), \pi_h(s_h)),$$

for any $(s_h, x_h^{(i)}, b_h, h, i) \in \mathcal{S} \times \mathcal{X}^{(i)} \times \mathcal{B} \times [H-1] \times [N]$. For step H, we define $Q_H^{(\pi,\nu),(i)}(s_H, x_H^{(i)}) = u_H^{(i)}(s_H, x_H^{(i)})$ and $V_H^{(\pi,\nu),(i)}(s_H) = Q_H^{(\pi,\nu),(i)}(s_H, \nu_H^{(i)}(s_H))$. By the definition, all these functions take value between 0 and H.

2.2. The Goal of MEE: Social Welfare Maximization with Competitive Equilibrium

Now we specify the goal for both social planner and agents in an MEE, that is, the agents aim to achieve *competitive equilibrium* at each step and the planner aims to *maximize the social welfare* which is the sum of utilities of all agents. We first study the optimal policy for the agents and the planner respectively, and after we define the joint optimality for planner-agents policy pair (π, ν). The joint optimality can be characterized by a fixed-point formulation, which allows us to define the suboptimality for any policy pair.

One-Step Competitive Equilibrium. The agents' optimal policy $\nu^\star$ is defined as the one giving *competitive equilibrium* with respect to the utility functions $\{u^{(i)}\}_{i\in[N]}$ at each step h. To this end, we first define a competitive equilibrium (Mas-Colell et al., 1995; Guo et al., 2021) as follows, which is adapted to the Markovian exchange economy.

Definition 2.1 (Competitive Equilibrium)**.** A competitive equilibrium (CE) at state $s = (c, e^{(1)}, \cdots, e^{(N)}) \in \mathcal{S}$ is an allocation and price-vector pair $(x^{(1),\star}, \cdots, x^{(N),\star}, p^\star) \in \mathcal{A}$ such that (i) the allocation is feasible and (ii) all agents maximize their utilities under the budget induced by price $p^\star$. In other words, following two conditions hold,

$$\sum_{i\in[N]} x_j^{(i),\star} \leqslant \sum_{i\in[N]} e_j^{(i)}, \ \forall j \in [L], \quad (2)$$
$$x^{(i),\star} \in \operatorname*{arg\,max}_{(x^{(i)})^\top p^\star \leq (e^{(i)})^\top p^\star} u^{(i)}(s, x^{(i)}), \ \forall i \in [N]. \quad (3)$$

For simplicity, we denote any competitive equilibrium allocation and price pair at state s with respect to $\{u^{(i)}\}_{i\in[N]}$ as $(x^{(1),\star}, \cdots, x^{(N),\star}, p^\star) \in \mathtt{CE}(\{u^{(i)}(s,\cdot)\}_{i\in[N]})$. Based on Definition 2.1, we define the agents' optimal policy as the one that outputs CE pairs at each time step.

Definition 2.2 (Optimal Policy of Agent)**.** The agents' optimal policy $\nu^\star$ is the policy in $\mathbf{\Pi}$ such that for any $(h, s_h) \in [H] \times \mathcal{S}$, $\nu_h^\star(s_h) = (\nu_h^{\star,(1)}(s_h), \cdots, \nu_h^{\star,(N)}(s_h), \nu_h^{\star,\mathbf{P}}(s_h))$ satisfies $\nu_h^\star(s_h) \in \mathrm{CE}(\{u_h^{(i)}(s_h, \cdot)\}_{i\in[N]})$.

Under certain assumptions (Mas-Colell et al., 1995; Guo et al., 2021) on the utility functions $\{u_h^{(i)}\}_{i\in[N],h\in[H]}$, the competitive equilibrium exists. To measure the suboptimality of a given policy ν, we further define the best responce policy of ν that reallocates among agents given the price system of ν to achieve competitive equilibrium.

Definition 2.3 (Best Responce of Agent Policy)**.** Given any agents' policy $\nu \in \mathbf{\Pi}$, the best responce agents' policy $\nu^\star(\nu)$ is the one in $\mathbf{\Pi}$ such that for any $(h, s_h) \in [H] \times \mathcal{S}$, $\nu_h^\star(\nu)(s_h) = (\nu_h^{\star,(1)}(\nu), \cdots, \nu_h^{\star,(N)}(\nu), \nu_h^{\star,\mathbf{P}}(\nu))(s_h)$ satisfies $\nu_h^{\star,\mathbf{P}}(\nu)(s_h) = \nu_h^{\mathbf{P}}(s_h)$ and $\nu_h^{\star,(i)}(\nu)(s_h) \in$

$$\operatorname*{arg\,max}_{x_h^{(i)}\in\mathcal{X}^{(i)}:(x_h^{(i)})^\top \nu_h^{\mathbf{P}}(s_h)\le (e_h^{(i)})^\top \nu_h^{\mathbf{P}}(s_h)} u_h^{(i)}(s_h, x_h^{(i)}). \quad (4)$$

The existence of $\nu^\star(\nu)$ is guaranteed by the theorem of the maximum, see Theorem A.2.21 of (Jehle, 2001). With the best responce agent policy $\nu^\star(\nu)$, we can measure the suboptimality of ν by comparing the value functions induced by ν and $\nu^\star(\nu)$, which also gives a fixed-point formulation of the agents' optimal policy. We conclude this property in the following theorem whose proof is in Appendix E.1.

Theorem 2.4. *For any policy pair (π, ν) satisfying the resource constraints, i.e., for any $(j, h, s_h) \in [L] \times [H] \times \mathcal{S}$,*

$$\sum_{i=1}^N (\nu_h^{(i)}(s_h))_j \le \sum_{i=1}^N (e_h^{(i)})_j,$$

the following two conclusions hold. (i) For any step $h \in [H]$ and state $s_h \in \mathcal{S}$, we have that

$$V_h^{(\pi,\nu),(i)}(s_h) \le V_h^{(\pi,\nu^\star(\nu)),(i)}(s_h). \quad (5)$$

(ii) If the equality $V_1^{(\pi,\nu^\star(\nu)),(i)}(s_1) = V_1^{(\pi,\nu),(i)}(s_1)$ holds for any $s_1 \in \mathcal{S}$, then for any $h \in [H]$ and $s_h \in \mathcal{S}$, $\nu_h(s_h)$ is a competitive equilibrium with respect to $\{u_h^{(i)}(s_h, \cdot)\}_{i\in[N]}$.

Theorem 2.4 tells that we end up higher values when substituting ν by $\nu^\star(\nu)$. Whenever the equality holds, the agent policy ν is optimal. Motivated by this fixed-point characterization of $\nu^\star$, we define the suboptimality of agents' policy to be $\mathrm{SubOpt}_{\mathrm{A}}^{(i)}(\pi, \nu, s_1)$ for each agent $i \in [N]$ as

$$V_1^{(\pi,\nu^\star(\nu)),(i)}(s_1) - V_1^{(\pi,\nu),(i)}(s_1). \quad (6)$$

Social Welfare Maximization. For given agents' policy ν, we define the planner's optimal $\pi^\star(\nu)$ to be the one in $\mathbf{N}$ that maximizes the social walfare $\sum_{i=1}^N V_1^{(\nu,\pi),(i)}(s_1)$, i.e., the sum of utility functions over agents and time steps.

Definition 2.5 (Optimal Policy of Planner)**.** The planner's optimal policy $\pi^\star(\nu)$ given agents' policy ν is the one in $\mathbf{N}$ such that for any $(h, s_h) \in [H] \times \mathcal{S}$, $\pi_h^\star(s_h)$ belongs to

$$\operatorname*{arg\,max}_{b_h\in\mathcal{B}} \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{(\pi^\star(\nu),\nu),(i)}(s') P_h(\mathrm{d}s'|s_h, b_h). \quad (7)$$

We note that $V_{h+1}^{(\pi^\star(\nu),\nu),(i)}$ only depends on $\pi_j^\star(\nu)$ for $j > h$ and thus $\pi^\star(\nu)$ is well-defined. Given any policy pair (π, ν), we can measure the suboptimality of π with respect to $\pi^\star(\nu)$ by comparing the social walfare induced by π and $\pi^\star(\nu)$. We show this result by the following theorem proven in Appendix E.2.

Theorem 2.6. *For any policy pair (π, ν), the following two conclusions hold. (i) For any step $h \in [H]$ and state $s_h \in \mathcal{S}$,*

$$\sum_{i=1}^N V_h^{(\pi,\nu),(i)}(s_h) \le \sum_{i=1}^N V_h^{(\pi^\star(\nu),\nu),(i)}(s_h). \quad (8)$$

(ii) Furthermore, if the equality $\sum_{i=1}^N V_1^{(\pi^\star(\nu),\nu),(i)}(s_1) = \sum_{i=1}^N V_1^{(\pi,\nu),(i)}(s_1)$ holds for any $s_1 \in \mathcal{S}$, then for any $h \in [H]$ and $s_h \in \mathcal{S}$ we have that

$$\pi_h(s_h) \in \operatorname*{arg\,max}_{b_h\in\mathcal{B}} \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{(\pi^\star(\nu),\nu),(i)}(s') P_h(\mathrm{d}s'|s_h, b_h).$$

Parallel to Theorem 2.4, Theorem 2.6 shows that we end up higher values when substituting π by $\pi^\star(\nu)$. Whenever the equality holds, the planner policy π is optimal given ν. Motivated by the fixed-point formulation of $\pi^\star(\nu)$, we define the suboptimality of planner's policy $\mathrm{SubOpt}_{\mathrm{P}}(\pi, \nu, s_1)$ as

$$\sum_{i=1}^N V_1^{(\pi^\star(\nu),\nu),(i)}(s_1) - V_1^{(\pi,\nu),(i)}(s_1). \quad (9)$$

Joint Optimality. Now we define the jointly optimal policy for the planner and the agents as $(\pi^\star(\nu^\star), \nu^\star)$, where $\nu^\star$ and $\pi^\star(\nu^\star)$ satisfies Definition 2.2 and 2.5 repectively, i.e., the agents find one-step CE and the planner maximizes the social walfare induced by the agents' CE policy. Based on the suboptimality (6) and (9) for agents and planner, we further define the suboptimality $\mathrm{SubOpt}(\pi, \nu, s_1)$ for any planner-agents policy pair (π, ν) as the following sum,

$$\sum_{i=1}^N \mathrm{SubOpt}_{\mathrm{A}}^{(i)}(\pi, \nu, s_1) + \mathrm{SubOpt}_{\mathrm{P}}(\pi, \nu^\star(\nu), s_1). \quad (10)$$

Plugging in the expression of suboptimalities (6) and (9), the suboptimality (10) is equivalent to the following expression,

$$\sum_{i=1}^N V_1^{(\pi^\dagger(\nu),\nu^\star(\nu)),(i)}(s_1) - V_1^{(\pi,\nu),(i)}(s_1), \quad (11)$$

where for simplicity, we denote $\pi^\dagger(\nu) := \pi^\star(\nu^\star(\nu))$ the optimal planner policy given the best responce agent policy of ν. We keep to this notation in the sequel. The following theorem, a corollary of Theorem 2.4 and 2.6, shows that the joint optimality is equivalent to that (11) vanishes.

Theorem 2.7 (Fixed-Point Characterization of Joint Optimality)**.** *A planner-agents policy pair (π, ν) is jointly optimal if and only if* $\mathrm{SubOpt}(\pi, \nu, s_1)$ (11) *equals to zero.*

2.3. Fair Division Property

Achieving competitive equilibrium among agents is also related to the notion of *fair division* mechanism which requires sharing incentive (SI) and Pareto-efficiency (PE) (Varian, 1973; Budish et al., 2017; Babaioff et al., 2019; 2021; Guo et al., 2021). An allocation $x_h \in \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(n)}$ satisfies SI at step h and state $s_h = (c_h, e_h)$ if the utility the i-th agent receives is at least as much as its utility when using its endowment, i.e. $u_h^{(i)}(s_h, x_h^{(i)}) \geq u_h^{(i)}(s_h, e_h^{(i)})$. This implies that all agents have the incentive to participate in this division mechanism. Besides, a feasible allocation x_h is PE at step h and state $s_h = (c_h, e_h)$ if the utility of one agent can be increased only by decreasing the utility of others. Formally, allocation x_h is said to dominate another allocation $\widetilde{x}_h$ given state s_h, if $u_h^{(i)}(s_h, x_h^{(i)}) \geq u_h^{(j)}(s_h, \widetilde{x}_h^{(i)})$ for all $i \in [N]$ and there exists some $j \in [N]$ such that $u_h^{(i)}(s_h, x_h^{(i)}) > u_h^{(j)}(s_h, \widetilde{x}_h^{(i)})$. An allocation x_h is Pareto-efficient given state s_h if it is not dominated by any other allocations. We denote the set of Pareto-efficient allocations at step h and state s_h by $\mathcal{PE}(s_h, h)$.

To characterize the fair division property when finding the optimal policy of agents, we further introduce corresponding loss functions. We first define the SI loss ℓ_h^{SI} for any agents' policy ν at step h and state s_h as the sum, over all agents, of how much they are worse off than their endowment utilities, i.e., we define $\ell_h^{\mathrm{SI}}(\nu, s_h)$ as

$$\sum_{i=1}^{N} (u_h^{(i)}(s_h, e_h^{(i)}) - u_h^{(i)}(s_h, \nu_h^{(i)}(s_h)))^+. \tag{12}$$

Then we define the PE loss ℓ_h^{PE} for ν at step h and state s_h as the minimal sum, over all agents, of how much they are worse off than PE allocations, i.e., we define $\ell_h^{\mathrm{PE}}(\nu, s_h)$ as

$$\inf_{x \in \mathcal{PE}(s_h, h)} \sum_{i=1}^{N} (u_h^{(i)}(s_h, x^{(i)}) - u_h^{(i)}(s_h, \nu_h^{(i)}(s_h)))^+. \tag{13}$$

Finally, we define the FD loss ℓ_h^{FD} for policy ν at step h as the maximum of SI loss ℓ_h^{SI} and PE loss ℓ_h^{PE}, i.e.,

$$\ell_h^{\mathrm{FD}}(\nu, s_h) = \max\left\{\ell_h^{\mathrm{PE}}(\nu, s_h), \ell_h^{\mathrm{SI}}(\nu, s_h)\right\}. \tag{14}$$

2.4. General Function Approximation and CE Oracle

In this paper, we apply MARL-style approaches to solve MEE in both online and offline settings with *general function approximations*. Specifically, we consider two function classes $\mathcal{U}$ and $\mathcal{P}$ to represent the utility functions $\{u_h^{(i)}\}_{(i,h)\in[N]\times[H]}$ and the transition kernels $\{P_h\}_{h\in[H]}$ respectively. We make the following realizability assumptions (Uehara & Sun, 2021; Xie et al., 2021) on them.

Assumption 2.8 (Realizability)**.** Without loss of generality, we assume that $\mathcal{X}^{(i)}$'s are the same for all $i \in [N]$. Then we assume that utility function $u_h^{(i)} \in \mathcal{U}$ and transition $P_h \in \mathcal{P}$ holds for any $(i, h) \in [N] \times [H]$.

Besides, we assume that for each set of $\{u^{(i)}\}_{i\in[N]}$ in $\mathcal{U}$, there exists a CE oracle $\mathtt{CE}(\{u^{(i)}(s,\cdot)\}_{i\in[N]})$ for any $s \in \mathcal{S}$ that returns CE allocation-price vector pair. This can be realized efficiently via methods introduced in Varian & Varian (1992); Zhang (2011); Zahedi et al. (2018).

3. Online Learning Algorithm

3.1. Setup and Learning Objective

Online Learning Protocol. We study online episodic setting where an online learning algorithm plays an MEE for K episodes. At the beginning of the k-th episode, the algorithm determines the planner's and agents' policy pair (π^k, ν^k), and an initial state s_1^k is chosen by the environment. At each time step $h \in [H]$, the agents and the planner observe state $s_h^k \in \mathcal{S}$ and pick their own actions $a_h^k = \nu_h^k(s_h^k)$ and $b_h^k = \pi_h^k(s_h^k)$. Subsequently, the environment transits to the next state $s_{h+1}^k \sim P_h(\cdot|s_h^k, b_h^k)$ and they observe the utilities $\{u_h^{k,(i)}\}_{i\in[N]}$ with $u_h^{k,(i)} = u_h^{(i)}(s_h^k, x_h^{k,(i)})$.

Learning Objective. Based on the definition of suboptimality (11) for any policy pair (π, ν), we define the following online regret with respect to achieving joint optimality.

Definition 3.1 (Online Regret for Joint Optimality)**.** Let (π^k, ν^k) be the policy pair executed by any online learning algorithm in the k-th episode. After a total of K episodes, the online regret for joint optimality is defined as

$$\mathrm{Regret}_{\mathrm{CE,SWM}}(K) = \sum_{k=1}^{K} \mathrm{SubOpt}(\pi^k, \nu^k, s_1^k). \tag{15}$$

Moreover, we also define the online regret with respect to achieving fair division based on the notion of FD loss (14).

Definition 3.2 (Online Regret for Fair Division Property)**.** Let (π^k, ν^k) be the policy pair executed by any online algorithm in the k-th episode. After a total of K episodes, the

the online regret for fair division property is defined as

$$\mathrm{Regret}_{\mathrm{FD}}(K) = \sum_{k=1}^{K} \mathbb{E}_{\pi^k}\left[\sum_{h=1}^{H} \ell_h^{\mathrm{FD}}(\nu^k, s_h)\right]. \quad (16)$$

Each summand in (16) reflects the expected total FD loss of ν^k along the trajectories induced by π^k. We remark that $\mathrm{Regret}_{\mathrm{FD}}(K)$ is an extension of the FD loss defined in (Guo et al., 2021) to sequantial settings. Our goal in the online setting is to design algorithms with both regrets sublinear in K , and polynomial in d and H, where d is some dimension of the function class used by the algorithm.

3.2. Algorithm: Model-based Optimistic Online Learning for MEE

We propose **M**odel-based **O**ptimistic online **L**earning for **M**EE (MOLM, Algorithm 1) to learn the joint optimal policy for planner and agents in the online setting, which involves a model estimation step and an optimistic planning step.

Model Estimation Step (Line 3). At the beginning of k-th episode, we construct confidence sets for the utility $u_h^{(i)}$ and the transition P_h using data collected before the k-th episode, inspired by Russo & Van Roy (2013); Ayoub et al. (2020); Cai et al. (2020b). For utility $u_h^{(i)}$, we let $u_h^{k,(i)}$ minimize the empirical mean squared error in $\mathcal{U}$ and let confidence set $\mathcal{U}_h^{k,(i)}$ consist of all the utility functions in $\mathcal{U}$ with empirical mean squared discrepency from $u_h^{k,(i)}$ less than a given threshold $\beta^{(1)}$. For transition P_h, we similarly construct the confidence set $\mathcal{P}_h^k$ via value-targeted regression (Ayoub et al., 2020; Cai et al., 2020b). Given value function estimators $\{V_{h+1}^{\tau,(i)}\}_{\tau=1}^{k-1}$, we let P_h^k minimize the empirical mean squared error in predicting the value of future social welfare $\sum_{i=1}^N V_{h+1}^{\tau,(i)}$ given s_h^τ, b_h^τ, and the confidence set $\mathcal{P}_h^k$ constains all transitions in $\mathcal{P}$ that make similar predictions to P_h^k with empirical mean squared discrepency less than another threshold $\beta^{(2)}$. Details of the model estimation step are concluded in Algorithm 3 in Appendix C.

Optimistic Planning Step (Line 4 to Line 8). Then using $\mathcal{U}_h^{k,(i)}$ and $\mathcal{P}_h^k$, MOLM performs optimistic planning according to (1), Definition 2.2 and 2.5 to obtain (π^k, ν^k) which is executed in the k-th episode. Intuitively, we first solve the sub-problem of one-step CE for the agents by choosing the optimal agents' policy with respect to the estimated optimistic utilities. After, we cast the sub-problem of social welfare maximization for the planner as finding the optimal policy in an Markov decision process whose reward is induced by the agents' utilities. Specifically, given the value function estimators $V_{h+1}^{k,(i)}$ at step $h+1$ with $V_{H+1}^{k,(i)}$ being zero functions, we first choose the most optimistic model estimator at step h from $\mathcal{U}_h^{k,(i)}$ and $\mathcal{P}_h^k$ respectively as

$$\widehat{u}_h^{k,(i)}(s, x^{(i)}) = \arg\max_{u \in \mathcal{U}_h^{(i),k}} u(s, x^{(i)}), \quad (17)$$

$$\widehat{P}_h^k(\cdot|s,b) = \arg\max_{P \in \mathcal{P}_h^k} \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{k,(i)}(s') P(\mathrm{d}s'|s,b), \quad (18)$$

for any $(i, s, x^{(i)}, b) \in [N] \times \mathcal{S} \times \mathcal{X}^{(i)} \times \mathcal{B}$. Then we choose the agents' policy $\nu_h^k = (\nu_h^{k,(1)}(s), \cdots, \nu_h^{k,(N)}(s), \nu_h^{k,\mathbf{p}}(s))$ so as to output CE pairs with respect to the estimated optimistic utility function $\widehat{u}_h^{k,(i)}$ by a CE oracle (Section 2.4),

$$\nu_h^k(s) = \mathtt{CE}(\{\widehat{u}_h^{k,(i)}(s,\cdot)\}_{i\in[N]}). \quad (19)$$

Meanwhile, we choose the planner policy π_h^k so as to maximize the estimated optimistic future social welfare,

$$\pi_h^k(s) = \arg\max_{b \in \mathcal{B}} \sum_{i=1}^N \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s') \widehat{P}_h^k(\mathrm{d}s'|s,b). \quad (20)$$

We note that $V_{h+1}^{k,(i)}$ can be seen as the state-value function of a finite-horizon MDP whose reward of state s and action b is given by $\widehat{u}_h^{k,(i)}(s, \nu_h^k(s), b)$. After that, the value function estimators at step h are updated accordingly, i.e.,

$$\begin{aligned} Q_h^{k,(i)}(s, x_h^{(i)}, b) &= \widehat{u}_h^{k,(i)}(s, x_h^{(i)}) + \\ &\quad \mathtt{Clip}_{[0,H-h]}\left\{\int_{\mathcal{S}} V_{h+1}^{k,(i)}(s') \widehat{P}_h^k(\mathrm{d}s'|s,b)\right\}, \quad (21) \\ V_h^{k,(i)}(s) &= Q_h^{k,(i)}\big(s, \nu_h^{k,(i)}(s), \pi_h^k(s)\big), \end{aligned}$$

for any $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, where we clip the second term in $Q_h^{k,(i)}$ between 0 and $H - h$ due to the assumption that utility functions fall in the range $[0, 1]$. Finally, MOLM executes the joint policy (π_h^k, ν_h^k) and collects the data for the k-th episode according to the protocol in Section 3.1

3.3. Main Theoretical Results for Online Learning

Our main theoretical results are upper bounds on the two online regrets (15) and (16) incurred by Algorithm 1. For the analysis, we introduce the notion of *eluder dimension* which is firstly proposed by (Russo & Van Roy, 2013).

Definition 3.3 (Eluder Dimension)**.** Let $\mathcal{Z}$ be a set of real-valued functions on $\mathcal{X}$. For any $\varepsilon > 0$ and $\tau \in [K]$, we say that $x_\tau \in \mathcal{X}$ is $(\mathcal{Z}, \varepsilon)$-independent of $x_1, \cdots, x_{\tau-1} \in \mathcal{X}$ if there exists $f_1, f_2 \in \mathcal{Z}$ such that both $\sum_{j=1}^{\tau-1} |f_1(x_j) - f_2(x_j)|^2 \le \varepsilon^2$ and $|f_1(x_\tau) - f_2(x_\tau)| > \varepsilon$ hold. The *eluder dimension* of $\mathcal{Z}$ at scale ε, denoted by $\dim_{\mathbb{E}}(\mathcal{Z}, \varepsilon)$, is then defined as the length of the longest sequence $\{x_j\}_{j=1}^{\tau}$ such that x_j is $(\mathcal{Z}, \varepsilon)$-independent of $\{x_i\}_{i=1}^{j-1}$ for any $j \in [\tau]$.

We refer to Russo & Van Roy (2013) for a detailed discussion of the eluder dimension. For simplicity, we define

Algorithm 1 Model-based Optimistic Online Learning for Markov Exchange Economy (MOLM)

Input: Optimism parameters $\beta^{(1)}$ and $\beta^{(2)}$. Function classes $\mathcal{U}$ and $\mathcal{P}$.

Initialize dataset $\mathcal{D}_h^0 = \emptyset$ for all $h \in [H]$. Set $V_{H+1}^{k,(i)}(\cdot) = 0$ for all $(i,k) \in [N] \times [K]$.
for $k = 1$ to K **do**
  $\{\mathcal{U}_h^{k,(i)}\}_{(h,i)\in[H]\times[N]}, \{\mathcal{P}_h^k\}_{h\in[H]} = \texttt{ME}(\mathcal{U}, \mathcal{P}, \{\mathcal{D}_h^k\}_{h\in[H]}, \beta^{(1)}, \beta^{(2)})$ // Model Estimation, Algorithm 3.
  $\pi^k, \nu^k, \{V_h^{k,(i)}\}_{(h,i)\in[H]\times[N]} = \texttt{OPL}(\{\mathcal{U}_h^{k,(i)}\}_{(h,i)\in[H]\times[N]}, \{\mathcal{P}_h^k\}_{h\in[H]})$. // Optimistic Planning, Algorithm 4.
  Observe initial state s_h^k of episode k.
  for $h = 1$ to H **do**
    Take actions $a_h^k = \nu_h^k(s_h^k)$ and $b_h^k = \pi_h^k(s_h^k)$. Observe the next state s_{h+1}^k and the utilities $u_h^{k,(i)}$.
    Update dataset $\mathcal{D}_h^k = \mathcal{D}_h^{k-1} \cup \{s_h^k, a_h^k, b_h^k, \{u_h^{k,(i)}\}_{i\in[N]}, \{V_h^{k,(i)}\}_{i\in[N]}\}$.
  end for
end for

$\mathcal{Z}_\mathcal{P}$ to be the class of mappings $z_P : \mathcal{S} \times \mathcal{B} \times \{f : \mathcal{S} \mapsto [0, HN]\} \mapsto [0,1], (s, b, f(\cdot)) \mapsto \int_\mathcal{S} f(s')P(\mathrm{d}s' \mid s, b)$, for any $P \in \mathcal{P}$. With these preparations, we define dimension $d = \max\{\dim_\mathbb{E}(\mathcal{U}, 1/K), \dim_\mathbb{E}(\mathcal{Z}_\mathcal{P}, 1/K)\}$ to characterize the complexity of function classes $\mathcal{U}$ and $\mathcal{P}$. The following theorem is the main theoretical results in the online setting. All the omitted proofs are in Appendix C.1 and F.

Theorem 3.4 (Regret of Algorithm 1)**.** *By setting parameters $\beta^{(1)}$ as $C_1 \log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)NHK^2/\delta)$ and $\beta^{(2)}$ as $C_2 H^2 N^2 \log(\mathcal{N}(1/(NHK), \mathcal{P}, \|\cdot\|_{\infty,1})NHK^2/\delta)$ for some absolute constants C_1 and C_2 in Algorithm 1, it holds with probability at least $1-\delta$ that the regret for joint optimality (15) and the regret for fair division property (16) of Algorithm 1 satisfies that*

$$\mathrm{Regret}_{\mathrm{CE,SWM}}(K) \le \mathcal{O}(\sqrt{dH^2(N^2\beta^{(1)} + \beta^{(2)})K}). \quad (22)$$

$$\mathrm{Regret}_{\mathrm{FD}}(K) \le \mathcal{O}(\sqrt{dH^2N^2\beta^{(1)}K}). \quad (23)$$

We show the exact expression of $\beta^{(1)}$ and $\beta^{(2)}$ in the proof of Theorem 3.4 in Appendix C.1. Theorem 3.4 indicates that the regret for joint optimality of Algorithm 1 is of order $\widetilde{\mathcal{O}}(\sqrt{dH^4N^2K})$, which shows that MOLM efficiently finds the jointly optimal policy approximately. Besides joint optimality, Algorithm 1 also achieves fair division among agents approximately as a byproduct, i.e., it approximately finds agents' policy which simultaneously achieves SI and PE in the online setting. We specialize Theorem 3.4 to tabular, linear, and kernel cases in Appendix D.

4. Offline Learning Algorithm

4.1. Setup and Learning Objective

Offline Learning Protocol. Now we study the offline setting where the learner only has access to an offline dataset $\mathcal{D} = \{(s_h^\tau, \{x_h^{\tau,(i)}\}_{i\in[N]}, b_h^\tau, \{u_h^{\tau,(i)}\}_{i\in[N]})\}_{(\tau,h)\in[K]\times[H]}$ which is generated as a prior by an economist in the MEE. We characterize the generation process of $\mathcal{D}$ by the following definition.

Definition 4.1 (Offline Data Generation)**.** The dataset $\mathcal{D}$ consists of K i.i.d. trajectories $\{\mathcal{D}^\tau\}_{\tau\in[K]}$, where each trajectory $\mathcal{D}^\tau = \{(s_h^\tau, \{x_h^{\tau,(i)}\}_{i\in[N]}, b_h^\tau, \{u_h^{\tau,(i)}\}_{i\in[N]})\}_{h\in[H]}$ is collected as a prior in the MEE. Specifically, for each $\tau \in [K]$, it holds that $s_{h+1} \sim P_h(\cdot \mid s_h^\tau, b_h^\tau)$, $u_h^{\tau,(i)} = u_h(s_h^\tau, x_h^{\tau,(i)})$.

Learning Objective. In offline learning, the goal is to design algorithm that outputs policy pair $(\widehat{\pi}, \widehat{\nu})$ which is joint optimal and achieves fair division. For being joint optimal, we measure the performance of $(\widehat{\pi}, \widehat{\nu})$ by $\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}, s_1)$ defined in (11). For achieving fair division, we adapt the FD loss defined in (14) to offline setting as follows.

$$\mathcal{L}_{\mathrm{FD}}(\pi, \nu) = \sum_{h=1}^H \mathbb{E}_{\rho_h}\left[\ell_h^{\mathrm{FD}}(\nu, s_h)\right], \quad (24)$$

where ρ_h is the visitation measure at step $h \in [H]$ that the dataset $\mathcal{D}$ obeys, i.e., $\rho_h(s, \{x^{(i)}\}_{i\in[N]}, b)$ is defined as

$$\mathbb{P}(s_h^\tau = s, \{x^{\tau,(i)}\}_{i\in[N]} = \{x_h^{(i)}\}_{i\in[N]}, b_h^\tau = b), \quad (25)$$

for any $\tau \in [K]$ in $\mathcal{D}$. We hope to design an algorithm that achieves suboptimality $\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}, s_1)$ and offline FD loss $\mathcal{L}_{\mathrm{FD}}(\widehat{\pi}, \widehat{\nu})$ decaying at a negative square root rate with respect to K.

4.2. Algorithm: Model-based Pessimistic Offline Learning for MEE

We propose **M**odel-based **P**essimistic offline **L**earning for **MEE** (MPLM, Algorithm 2) to learn the desired planner-agent policy pair, which involves a model estimation step and a pessimistic policy optimization step. We use $\widehat{V}_{h,(\widehat{P},\widehat{u})}^{(\pi,\widehat{\nu}),(i)}$ to denote the value function of policy pair $(\widehat{\pi}, \widehat{\nu})$ induced by the estimated utility function $\widehat{u} = \{\widehat{u}_h^{(i)}\}_{(h,i)\in[H]\times[N]}$ and the estimated transition $\widehat{P} = \{\widehat{P}_h\}_{h\in[H]}$ according to (1).

Algorithm 2 Model-based Pessimistic Online Learning for Markov Exchange Economy (MPLM)

Input: Pessimism Parameter ξ_1, ξ_2. Function classes $\mathcal{U}$ and $\mathcal{P}$.

1: **for** $h = 1$ to H **do**
2: Construct confidence sets $\{\mathcal{U}_{h,\xi_1}^{(i)}\}_{i\in[N]}$ and $\mathcal{P}_{h,\xi_2}$ according to (26), (27). // Model Estimation.
3: **end for**
4: **for** $h = 1$ to H **do**
5: Set $\widehat{u}_h^{(i)}(s, x^{(i)}) = \arg\min_{u\in\mathcal{U}_{h,\xi_1}^{(i)}} u(s, x^{(i)})$ and $\widehat{\nu}_h(s) = \mathtt{CE}(\{\widehat{u}_h^{(i)}(s,\cdot)\}_{i\in[N]})$, for any $(i, s, x^{(i)}) \in [N]\times\mathcal{S}\times\mathcal{X}^{(i)}$.
6: **end for**
7: Set $(\widehat{\pi}, \widehat{P}) = \arg\max_{\pi\in\Pi}\min_{\widehat{P}:\{\widehat{P}_h\in\mathcal{P}_{h,\xi_2},\forall h\in[H]\}} \cdot \sum_{i=1}^N \widehat{V}_{1,(\widehat{P},\widehat{u})}^{(\pi,\widehat{\nu}),(i)}(s_1)$. // Pessimistic Policy Optimization.

Output: Policy pair $(\widehat{\pi}, \widehat{\nu})$.

Model Estimation (Line 1 to 3). We first construct confidence sets for the utility $u_h^{(i)}$ and the transition P_h respectively. For $u_h^{(i)}$, we let the confidence set $\mathcal{U}_{h,\xi_1}^{(i)}$ consist of all functions in $\mathcal{U}$ whose empirical mean squared errors are less than a given threshold ξ_1, i.e., we set $\mathcal{U}_{h,\xi_1}^{(i)}$ as

$$\left\{u\in\mathcal{U} : \frac{1}{K}\sum_{\tau=1}^{K}\left(u_h^{\tau,(i)} - u(s_h^\tau, x_h^{\tau,(i)})\right)^2 \le \xi_1\right\}, \quad (26)$$

For P_h, we first obtain the maximum likelihood estimator $\widehat{P}_h^{\mathrm{MLE}}$ that maximizes the empirical likelihood function, i.e., $\widehat{P}_h^{\mathrm{MLE}} = \arg\max_{P\in\mathcal{P}}\sum_{\tau=1}^K \log P(s_{h+1}^\tau \mid s_h^\tau, b_h^\tau)$. Then we set the confidence set $\mathcal{P}_{h,\xi_2}$ to be transitions in $\mathcal{P}$ whose empirical mean squared TV-distance to $\widehat{P}_h^{\mathrm{MLE}}$ is less than a given threshold ξ_2, i.e., we set $\mathcal{P}_{h,\xi_2}$ as

$$\left\{P\in\mathcal{P} : \frac{1}{K}\sum_{\tau=1}^{K}\|(\widehat{P}_h^{\mathrm{MLE}} - P)(\cdot \mid s_h^\tau, b_h^\tau)\|_1^2 \le \xi_2\right\}. \quad (27)$$

Pessimistic Optimization (Line 4 to 8). With $\mathcal{U}_{h,\xi_1}^{(i)}$ and $\mathcal{P}_{h,\xi_2}$, MPLM then performs pessimistic policy optimization to find the policy pair $(\widehat{\pi}, \widehat{\nu})$ as its output. Parallel to the online setting, we first solve the sub-problem for agents via choosing the optimal agents' policy $\widehat{\nu}$ with respect to the estimated pessimistic utilities. After, we cast the sub-problem of social welfare maximization for the planner as an offline policy optimization problem in MDP with reward induced by the agents' utilities. Inspired by Uehara & Sun (2021), we solve this sub-problem by jointly optimizing over π and P such that the pessimistic social welfare estimator is maximized, which can be formulated as a minimax optimization problem. See Algorithm 2 for a detailed description.

4.3. Main Theoretical Results for Offline Learning

Our main theoretical results are upper bounds on the suboptimality (11) and offline FD loss (24) incurred by Algorithm 2. To guarantee provably efficient learning, we make certain assumptions on the coverage property of the dataset $\mathcal{D}$. Recall that the visitation measure ρ_h at step $h \in [H]$ the dataset $\mathcal{D}$ obeys is defined in (25). In parallel, we define the visitation measure $d_h^{(\pi,\nu)}(s, \{x^{(i)}\}_{i\in[N]}, b)$ at step $h \in [H]$ of any given joint policy (π, ν) as

$$\mathbb{P}(s_h = s, \{x^{(i)}\}_{i\in[N]} = \{x_h^{(i)}\}_{i\in[N]}, b_h = b \,|\, \pi, \nu). \quad (28)$$

Definition 4.2 (Distribution Shift)**.** We define the distribution shift coefficient between a given joint policy (π, ν) and the dataset visitation measure $\rho = \{\rho_h\}_{h\in[H]}$ as

$$C_\rho^{(\pi,\nu)} = \sup_{h\in[H]} \mathbb{E}_{\rho_h}\left(\frac{d_h^{(\pi,\nu)}(s, \{x^{(i)}\}_{i\in[N]}, b)}{\rho_h(s, \{x^{(i)}\}_{i\in[N]}, b)}\right)^2. \quad (29)$$

Assumption 4.3 (Partial Coverage)**.** We assume that the distribution shift between all the possible jointly optimal policy and the dataset visitation measure is finite, that is,

$$C_\rho^\star := \sup_{\nu\in\mathbf{N}^\star, \pi\in\mathbf{\Pi}^\star} C_\rho^{(\pi,\nu)} < \infty, \quad (30)$$

where $\mathbf{\Pi}^\star := \{\pi^\star(\nu) : \nu \in \mathbf{N}\}, \mathbf{N}^\star := \{\nu^\star(\nu) : \nu \in \mathbf{N}\}$.

Similar partial coverage assumptions are widely adopted in offline RL literature (Kidambi et al., 2020; Jin et al., 2021b), which is weaker than uniform coverage assumptions (Munos & Szepesvári, 2008; Chen & Jiang, 2019). The following two theorems are our main results in offline setting. All the omitted proofs are in Appendix C.2 and G.

Theorem 4.4 (Suboptimality of Algorithm 2)**.** *By setting the parameters ξ_1 as $C_1\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)\cdot NH/\delta)/K$ and ξ_2 as $C_2\log(\mathcal{N}_{[]}(1/K, \mathcal{P}, \|\cdot\|_{2,\infty})H/\delta)/K$ for some absolute constants C_1 and C_2 in Algorithm 2, it holds with probability at least $1-\delta$ that the suboptimality (11) and offline FD loss (24) of Algorithm 2 satisfies*

$$\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}) \le \mathcal{O}(\sqrt{H^4N^2\iota C_\rho^\star/K}), \quad (31)$$

$$\mathcal{L}_{\mathrm{FD}}(\widehat{\pi}, \widehat{\nu}) \le \mathcal{O}(\sqrt{H^2N^2\iota'/K}), \quad (32)$$

where $\iota = \log\mathcal{N}_{[]}(1/K^2, \mathcal{P}, \|\cdot\|_{1,\infty}) + \log\mathcal{N}(1/K^2, \mathcal{U}, \|\cdot\|_\infty) + \log(HN/\delta)$, $\iota' = \log(\mathcal{N}(1/K^2, \mathcal{U}, \|\cdot\|_\infty)\cdot NH/\delta)$ and $C_\rho^\star$ is defined in (30).

According to Theorem 4.4, the suboptimality of Algorithm 2 decays at a rate of $K^{-1/2}$, which shows that MPLM provably finds the jointly optimal policy approximately from dataset satisfying partial coverage assumptions. Parallel to the online setting, Theorem 4.4 also shows that Algorithm 2 achieves fair division among agents as a byproduct of achieving joint optimality in the offline setting, i.e., Algorithm 2 can output policy of agents which simultaneously achieves SI and PE in offline setting. We specialize Theorem 4.4 to tabular, linear, and kernel cases in Appendix D.

5. Experiments

To verify our theoretical results, we conduct an experiment for finding joint optimal policies for an MEE in online setting through our proposed MOLM algorithm. The codes are available on https://github.com/YSLIU627/RL-for-Markov-Exchange-Economy.

Settings. We focus on an MEE with a finite context space $|\mathcal{C}| = 10$, a finite planner's action space $|\mathcal{B}| = 5$, type of goods $L = 3$, a continuous agents' allocation space $[0, 1]^L$, number of agents $N = 3$, and horizon H=3. We randomly generate $P \in \mathbb{R}^{|\mathcal{C}|^2 \times |B|}$ for transition kernel and $A \in \mathbb{R}^{|H| \times |\mathcal{C}| \times N \times L}$ for utility functions by

$$u_h^{(i)}(c_h, x_h^{(i)}) = \sum_{j=1}^{L} A_{(h,c_h,i,j)} \cdot x_{h,j}^{(i)},$$

where we have let endowments $e^{(i)} = \delta_{ij}$ be time-invariant. Here $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

Practical Algorithm. We implement our MOLM algorithm for 100 epochs. In order to find CE given utility functions, we adopt the method of PRD (Brânzei et al., 2021) for 10 iterations. Also, for ease of implementation, we adopt the ucb-style optimism with uncertainty quantification

$$\Gamma_h(c, b) = \kappa |\mathcal{C}| H \sqrt{\frac{\log(H|\mathcal{C}||\mathcal{B}|/\delta)}{(N_h(c, b) \vee 1)}},$$

which is equivalent to finding the optimistic model in our confidence set (Yang & Wang, 2020). Here $N_h(c, b)$ denotes the times for visiting the pair (c, b), and we use $a \vee b$ to denote $\max\{a, b\}$. Finally, we adopt the Thompson sampling method (Guo et al., 2021) for optimistic utility estimation.

Experimental Results. We present the regret of MOLM (defined in (15), the sum of performance gap between MOLM's policy and joint optimal policy) in Figure 1, and we compare it to the linear regret induced by uniformly random policy, showing that MOLM *has the sublinear regret*, in correspondence to Theorem 3.4.

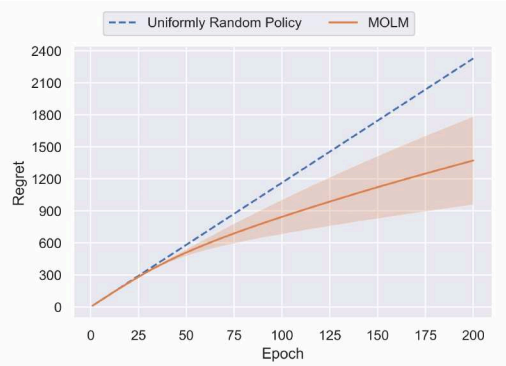

Figure 1. The regret of a uniformly random policy and the policy learned by MOLM. The results are averaged over 5 random seeds. We can see that the uniform random policy incurs a linear regret, while the policy learned by MOLM incurs a sublinear regret.

6. Discussions and Conclusions

In this work, we design provably efficient learning algorithms for online and offline MEE. Specifically, we propose MOLM/MPLM to learn agents' and planner's policies, respectively, achieving social welfare maximization and competitive equilibrium simultaneously, with fair division as a byproduct. To the best of our knowledge, we have presented the first provably efficient MARL-style algorithms for a bilevel economic system. Exploring and understanding learning methods in this general microeconomic setting promises to greatly extend the scope and range of applications of machine learning systems.

Acknowledgements

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Amazon, J.P. Morgan, and Two Sigma for their supports. We would like to thank Wenshuo Guo for valuable discussions.

References

Antos, A., Munos, R., and Szepesvári, C. Fitted q-iteration in continuous action-space mdps. 2007.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2009.

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

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

Babaioff, M., Kleinberg, R., and Slivkins, A. Multi-parameter mechanisms with implicit payment computation. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pp. 35–52, 2013.

Babaioff, M., Nisan, N., and Talgam-Cohen, I. Fair allocation through competitive equilibrium from generic incomes. In *FAT*, pp. 180, 2019.

Babaioff, M., Nisan, N., and Talgam-Cohen, I. Competitive equilibrium with indivisible goods and generic budgets. *Mathematics of Operations Research*, 46(1):382–403, 2021.

Bai, Y., Jin, C., Wang, H., and Xiong, C. Sample-efficient learning of stackelberg equilibria in general-sum games. *arXiv preprint arXiv:2102.11494*, 2021.

Balcan, M.-F. F., Sandholm, T., and Vitercik, E. Sample complexity of automated mechanism design. In *Advances in Neural Information Processing Systems*, pp. 2083–2091, 2016.

Başar, T. and Olsder, G. J. *Dynamic noncooperative game theory*. SIAM, 1998.

Bergemann, D. and Valimaki, J. Efficient dynamic auctions. 2006.

Blaug, M. The fundamental theorems of modern welfare economics, historically contemplated. *History of Political Economy*, 39(2):185–207, 2007.

Brânzei, S., Devanur, N., and Rabani, Y. Proportional dynamics in exchange economies. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 180–201, 2021.

Bucarey, V., Jean-Marie, A., Della Vecchia, E., and Ordóñez, F. On the value iteration method for dynamic strong stackelberg equilibria. In *ROADEF 2019-20ème congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision*, 2019.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Budish, E., Cachon, G. P., Kessler, J. B., and Othman, A. Course match: A large-scale implementation of approximate competitive equilibrium from equal incomes for combinatorial allocation. *Operations Research*, 65(2): 314–336, 2017.

Busoniu, L., Babuska, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

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

Cai, Q., Yang, Z., Szepesvari, C., and Wang, Z. Optimistic policy optimization with general function approximations. 2020b.

Cao, D. Recursive equilibrium in krusell and smith (1998). *Journal of Economic Theory*, 186:104978, 2020.

Chatrapati, K. S., Rekha, J. U., and Babu, A. V. Recursive competitive equilibrium approach for dynamic load balancing a distributed system. In *International Conference on Distributed Computing and Internet Technology*, pp. 162–174. Springer, 2011.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Cohen, K. J. and Cyert, R. M. Theory of the firm; resource allocation in a market economy. Technical report, Prentice-Hall, 1965.

Debreu, G. Existence of competitive equilibrium. *Handbook of mathematical economics*, 2:697–743, 1982.

Delimitrou, C. and Kozyrakis, C. Paragon: Qos-aware scheduling for heterogeneous datacenters. *ACM SIGPLAN Notices*, 48(4):77–88, 2013.

Dissanayake, I., Zhang, J., and Gu, B. Task division for team success in crowdsourcing contests: Resource allocation and alignment effects. *Journal of Management Information Systems*, 32(2):8–39, 2015.

Dudík, M., Haghtalab, N., Luo, H., Schapire, R. E., Syrgkanis, V., and Vaughan, J. W. Oracle-efficient online learning and auction design. In *2017 ieee 58th annual symposium on foundations of computer science (focs)*, pp. 528–539. IEEE, 2017.

Geer, S. A., van de Geer, S., and Williams, D. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Georgiadis, L., Neely, M. J., and Tassiulas, L. *Resource allocation and cross-layer control in wireless networks*. Now Publishers Inc, 2006.

Guo, W., Kandasamy, K., Gonzalez, J. E., Jordan, M. I., and Stoica, I. Online learning of competitive equilibria in exchange economies. *arXiv preprint arXiv:2106.06616*, 2021.

Hernandez-Leal, P., Kartal, B., and Taylor, M. E. Is multiagent deep reinforcement learning the answer or the question? a brief survey. *learning*, 21:22, 2018.

Hernandez-Leal, P., Kartal, B., and Taylor, M. E. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A. D., Katz, R. H., Shenker, S., and Stoica, I. Mesos: A platform for fine-grained resource sharing in the data center. In *NSDI*, volume 11, pp. 22–22, 2011.

Hussain, H., Malik, S. U. R., Hameed, A., Khan, S. U., Bickler, G., Min-Allah, N., Qureshi, M. B., Zhang, L., Yongji, W., Ghani, N., et al. A survey on resource allocation in high performance distributed computing systems. *Parallel Computing*, 39(11):709–736, 2013.

Jehle, G. A. *Advanced microeconomic theory*. Pearson Education India, 2001.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Annual Conference on Learning Theory*, 2019.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021a.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.

Kahn, A. J. *Theory and practice of social planning*. Russell Sage Foundation, 1969.

Kakade, S. M., Lobel, I., and Nazerzadeh, H. An optimal dynamic mechanism for multi-armed bandit processes. *arXiv preprint arXiv:1001.4598*, 2010.

Kandasamy, K., Gonzalez, J. E., Jordan, M. I., and Stoica, I. Mechanism design with bandit feedback. *arXiv preprint arXiv:2004.08924*, 2020.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

Kutschinski, E., Uthmann, T., and Polani, D. Learning competitive pricing strategies by multi-agent reinforcement learning. *Journal of Economic Dynamics and Control*, 27 (11-12):2207–2218, 2003.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Lussange, J., Lazarevich, I., Bourgeois-Gironde, S., Palminteri, S., and Gutkin, B. Modelling stock markets by multi-agent reinforcement learning. *Computational Economics*, 57(1):113–147, 2021.

Mankiw, N. G., Weinzierl, M., and Yagan, D. Optimal taxation in theory and practice. *Journal of Economic Perspectives*, 23(4):147–74, 2009.

Mannion, P., Mason, K., Devlin, S., Duggan, J., and Howley, E. Dynamic economic emissions dispatch optimisation using multi-agent reinforcement learning. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2016)*, 2016.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

Mehra, R. Recursive competitive equilibrium, 2006.

Min, Y., Wang, T., Xu, R., Wang, Z., Jordan, M. I., and Yang, Z. Learn to match with no regret: Reinforcement learning in markov matching markets. *arXiv preprint arXiv:2203.03684*, 2022.

Mirrlees, J. A. Optimal tax theory: A synthesis. *Journal of public Economics*, 6(4):327–358, 1976.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

OroojlooyJadid, A. and Hajinezhad, D. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

Rajaraman, N., Yang, L. F., Jiao, J., and Ramachandran, K. Toward the fundamental limits of imitation learning. *arXiv preprint arXiv:2009.05990*, 2020.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Rauch, D. E. and Schleicher, D. Like uber, but for local government law: the future of local regulation of the sharing economy. *Ohio St. LJ*, 76:901, 2015.

Ross, S. and Bagnell, J. A. Agnostic system identification for model-based reinforcement learning. *arXiv preprint arXiv:1203.1007*, 2012.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pp. 2256–2264. Citeseer, 2013.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Rzadca, K., Findeisen, P., Swiderski, J., Zych, P., Broniek, P., Kusmierek, J., Nowak, P., Strack, B., Witusowski, P., Hand, S., et al. Autopilot: workload autoscaling at google. In *Proceedings of the Fifteenth European Conference on Computer Systems*, pp. 1–16, 2020.

Salyer, K. D. Interpreting a stochastic monetary growth model as a modified social planner's problem. *Journal of Economic Dynamics and Control*, 20(4):681–689, 1996.

Sen, B. A gentle introduction to empirical process theory and applications, 2018.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Tiwari, M., Groves, T., and Cosman, P. C. Competitive equilibrium bitrate allocation for multiple video streams. *IEEE Transactions on Image Processing*, 19(4):1009–1021, 2009.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Varian, H. R. Equity, envy, and efficiency. 1973.

Varian, H. R. and Varian, H. R. *Microeconomic analysis*, volume 3. Norton New York, 1992.

Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., et al. Apache hadoop yarn: Yet another resource negotiator. In *Proceedings of the 4th annual Symposium on Cloud Computing*, pp. 1–16, 2013.

Venkataraman, S., Yang, Z., Franklin, M., Recht, B., and Stoica, I. Ernest: Efficient performance prediction for large-scale advanced analytics. In *13th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 16)*, pp. 363–378, 2016.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Wolski, R., Plank, J. S., Brevik, J., and Bryan, T. Analyzing market-based resource allocation strategies for the computational grid. *The International Journal of High Performance Computing Applications*, 15(3):258–281, 2001.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756, 2020.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

Zahedi, S. M., Llull, Q., and Lee, B. C. Amdahl's law in the datacenter era: A market for fair processor allocation. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 1–14. IEEE, 2018.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, pp. 12287–12297. PMLR, 2021.

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

Zhang, L. Proportional response dynamics in the fisher market. *Theoretical Computer Science*, 412(24):2691–2698, 2011.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020.

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., and Socher, R. The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., and Socher, R. The ai economist: Optimal economic policy design via two-level deep reinforcement learning. *arXiv preprint arXiv:2108.02755*, 2021.

Zhong, H., Yang, Z., Wang, Z., and Jordan, M. I. Can reinforcement learning find stackelberg-nash equilibria in general-sum markov games with myopic followers? *arXiv preprint arXiv:2112.13521*, 2021.

A. Notations

Throughout this paper, we denote by $[N] = \{1, ..., N\}$. We also denote by $\Delta(\mathcal{X})$ the probability space on set $\mathcal{X}$. We denote by $a = \mathcal{O}(b)$ if there exists an absolute constant c such that $a \leq cb$ when a and b are both large enough. We use $\widetilde{\mathcal{O}}(\cdot)$ to hide the constants term and logarithmic terms in $\mathcal{O}(\cdot)$. We use $\texttt{Clip}_{[a,b]}(c)$ to represent $\min\{\max\{a, c\}, b\}$ for real numbers a, b, and c. We use $\{a\}^+$ to represent $\max\{a, 0\}$ for real number a. Given a function class $\mathcal{F}$, we denote by $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$ the ϵ-covering number of $\mathcal{F}$ by $\|\cdot\|$ norm, and we denote by $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ the ϵ-bracket number of $\mathcal{F}$ by $\|\cdot\|$ norm. For function class $\mathcal{P} : \mathcal{S} \times \mathcal{B} \mapsto \Delta(\mathcal{S})$, we denote by $\|P\|_{1,\infty} = \sup_{s,b} \int_{\mathcal{S}} |P(s'|s, b)| \mathrm{d}s'$ for any $P \in \mathcal{P}$. We define that $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. For a distribution ρ, we use $\mathbb{E}_\rho[\cdot]$ and $\mathbb{V}_\rho[\cdot]$ to denote the expectation and the variance taken with respect to ρ, respectively.

General Notation	Explanation
$s_h = (c_h, e_h^{(1)}, \cdots, e_h^{(N)}) \in \mathcal{S}$	state at step h, $c_h \in \mathcal{C}$ is context, $e_h^{(i)} \in \mathcal{E}$ is endowments of the i^{th} agent
$a_h = (x_h^{(1)}, \cdots, x_h^{(N)}, p_h) \in \mathcal{A}$	action of agents at step h, $x_h^{(i)} \in \mathcal{X}^{(i)}$ is allocation of the i^{th} agent, $p_h \in [0, 1]^L$ is price
$b_h \in \mathcal{B}$	action of planner at step h
$u_h^{(i)}$	utility function of the i^{th} agent at step h
P_h	transition kernel at step h
$\nu = \{\nu_h\}_{h\in[H]}$	agents' policy, $\nu_h = (\nu_h^{(1)}, \cdots, \nu_h^{(N)}, \nu_h^{\mathbf{p}})$
$\nu^\star = \{\nu_h^\star\}_{h\in[H]}$	optimal agents' policy $\nu_h^\star = (\nu_h^{\star,(1)}, \cdots, \nu_h^{\star,(N)}, \nu_h^{\star,\mathbf{p}})$
$\nu^\star(\nu) = \{\nu_h(\nu)\}_{h\in[H]}$	best responce agents' policy of ν $\nu_h^\star(\nu) = (\nu_h^{\star,(1)}(\nu), \cdots, \nu_h^{\star,(N)}(\nu), \nu_h^{\star,\mathbf{p}}(\nu))$
$\pi = \{\pi_h\}_{h\in[H]}$	planner's policy
$\pi^\star(\nu) = \{\pi_h^\star(\nu)\}_{h\in[H]}$	optimal planner's policy given agents' policy ν (Definition 2.5)
$\pi^\dagger(\nu) = \{\pi_h^\dagger(\nu)\}_{h\in[H]}$	abbreviation of $\pi^\star(\nu^\star(\nu))$
$V_h^{(\pi,\nu),(i)}, Q_h^{(\pi,\nu),(i)}$	value functions of policy pair (π, ν) for the i^{th} agent at step h

Notations for Online Setting	Explanation
$s_h^k = (c_h^k, e_h^{k,(1)}, \cdots, e_h^{k,(N)}) \in \mathcal{S}$	state at step h of episode k
$a_h^k = (x_h^{k,(1)}, \cdots, x_h^{k,(N)}, p_h^k) \in \mathcal{A}$	action of agents at step h of episode k
$b_h^k \in \mathcal{B}$	action of planner at step h of episode k
$u_h^{k,(i)}$	utility of the i^{th} agent at step h of episode k
$\mathcal{U}_h^{k,(i)}$	confidence set of utility functions for the i^{th} agent at step h of episode k
$\mathcal{P}_h^k$	confidence set of transition kernels at step h of episode k
$\widehat{u}_h^{k,(i)}$	optimistic utility function estimator of the i^{th} agent at step h of episode k
$\widehat{P}_h^k$	optimistic transition estimator at step h of episode k
$\nu^k = \{\nu_h^k\}_{h\in[H]}$	agents' policy of episode k, $\nu_h^k = (\nu_h^{k,(1)}, \cdots, \nu_h^{k,(N)}, \nu_h^{k,\mathbf{p}})$
$\pi^k = \{\pi_h^k\}_{h\in[H]}$	planner's policy of episode k
$V_h^{k,(i)}, Q_h^{k,(i)}$	value function estimators at step h of episode k

Notation for Offline Setting	Explanation
$s_h^\tau = (c_h^\tau, e_h^{\tau,(1)}, \cdots, e_h^{\tau,(N)}) \in \mathcal{S}$	state at step h in dataset $\mathcal{D}$
$a_h^\tau = (x_h^{\tau,(1)}, \cdots, x_h^{\tau,(N)}, p_h^\tau) \in \mathcal{A}$	action of agents at step h in dataset $\mathcal{D}$
$b_h^\tau \in \mathcal{B}$	action of planner at step h in dataset $\mathcal{D}$
$u_h^{\tau,(i)}$	utility of the i^{th} agent in dataset $\mathcal{D}$
$\mathcal{U}_{h,\xi_1}^{(i)}$	confidence set of utility functions for the i^{th} agent at step h
$\mathcal{P}_{h,\xi_2}$	confidence set of transition kernels at step h
$\widehat{u}_h^{(i)}$	pessimistic utility function estimator of the i^{th} agent at step h
$\widehat{P}_h$	pessimistic transition estimator at step h
$\widehat{\nu} = \{\widehat{\nu}_h\}_{h\in[H]}$	estimated optimal agents' policy, $\widehat{\nu}_h = (\widehat{\nu}_h^{(1)}, \cdots, \widehat{\nu}_h^{(N)}, \widehat{\nu}_h^{\mathbf{p}})$
$\widehat{\pi} = \{\widehat{\pi}_h\}_{h\in[H]}$	estimated optimal planner's policy
$\widehat{V}_{h,(\widehat{P},\widehat{u})}^{(\pi,\widehat{\nu}),(i)}$	value function of $(\widehat{\pi},\widehat{\nu})$ induced by $\widehat{u} = \{\widehat{u}_h^{(i)}\}_{(h,i)\in[H]\times[N]}$ and $\widehat{P} = \{\widehat{P}_h\}_{h\in[H]}$

For the completeness of the paper, we provide the definitions of covering number and bracketing number as follows.

Definition A.1 (Covering Number by $\|\cdot\|$-norm)**.** Let $\mathcal{N}(\delta, \mathcal{G}, \|\cdot\|)$ be the smallest value of M for which there exist a subset $\{g_j^{\text{Cover}}\}_{j\in[M]} \subset \mathcal{G}$ such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in [M]$ such that

$$\|g_j^{\text{Cover}} - g\| \leq \delta.$$

Definition A.2 (Bracketing Number by $\|\cdot\|$-norm (Geer et al., 2000))**.** Let $\mathcal{N}_{[]}(\delta, \mathcal{G}, \|\cdot\|)$ be the smallest value of M for which there exist pairs of functions $\{[g_j^L, g_j^U]\}_{j\in[M]}$ such that $\|g_j^U - g_j^L\| \leq \delta$ for all $j \in [M]$, and for each $g \in \mathcal{G}$, there is a $j = j(g) \in [M]$ such that

$$g_j^L \leq g \leq g_j^U.$$

B. Related Work

We present detailed discussions on the related work in this section.

Machine learning for Economy. Our work adds to the vast body of existing literature on applying machine learning methods to solving various economical issues, where the utility functions for agents are not given a priori but learnable. For EE, Guo et al. (2021) propose the first online learning mechanism which adopts generalized linear function approximation and achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ online regret and online fair division loss. This theoretical result matches the the conclusion of our proposed mechanism, when MEE is specialized to EE and the function class of utility is chosen as generalized linear function. Aimed at analyzing the optimal allocation rules among agents, the automated mechanism design of revenue-maximizing combinatorial auctions has been widely studied with online learning methods (Bergemann & Valimaki, 2006; Kakade et al., 2010; Babaioff et al., 2013; Balcan et al., 2016; Dudík et al., 2017; Kandasamy et al., 2020). They analyze the online regret of the proposed mechanism even under the dynamic setting, while they do not consider bilevel economic systems and general function approximations for handling continuous state space as in this paper. Besides, several other works also adopt deep RL in multi-agent economic simulations, achieving empirical success (Zheng et al., 2021). Among them, Zheng et al. (2021) provide the first experimental MARL framework for the policy design of bilevel economic systems and obtain satisfactory results on the simulation baseline. However, the theory behind MARL methods is less studied. More recently, Min et al. (2022) apply MARL to study the problem of matching in a Markov matching market.

Exchange Economy. Our work is based on a rich line of aforementioned works in CE and fair division of EE (Cohen & Cyert, 1965; Debreu, 1982; Georgiadis et al., 2006; Zhang, 2011; Dissanayake et al., 2015). Under certain regularity conditions on utility functions, Debreu (1982) study the existence of CE of EE and Zhang (2011) propose a computationally efficient algorithm to compute CE of EE, both of which lay the foundation of our work. Besides, fair division of EE has been studied from both theoretical aspects (Varian, 1973; Budish et al., 2017; Babaioff et al., 2019; 2021) and practical points of view (Wolski et al., 2001; Vavilapalli et al., 2013; Zahedi et al., 2018). However, most previous works on EE assume the full knowledge of the utility functions of agents. As the initial attempt for learning unknown utility, some works (Zahedi et al., 2018; Le et al., 2020) assert some explicit but also restrictive assumptions on utility.

Moreover, Mehra (2006); Chatrapati et al. (2011); Cao (2020) study recursive CE (RCE) on dynamic EE, where agents exchange resources for multiple timesteps but no planner is involved. Cao (2020) study the existance of RCE under several restrictive assumptions on the transition kernel of dynamic EE. Different from MEE, dynamic EE neglects the co-adaptation between the agents and the planner and hence is not our interested bilevel system.

Social Planning Problem. Our work is also related to the social planning problem (SPP), a classic topic in welfare economics (Kahn, 1969; Salyer, 1996; Blaug, 2007). In SPP, a social planner who desires to maximize a predefined social welfare function can make all decisions in the economy. Different from EE, there is no price system in SPP, while the second social welfare theorem (Blaug, 2007) shows that any SPP can be decentralized to solving CE. Since SPP ignores the co-adaptation between the social planner and the agents in the economy, previous works on SPP can not solve or decentralize bilevel systems, such as MEE.

Multi-Agent Reinforcement Learning and Stackelberg Equilibrium. Our work is also related to a rich line of works in MARL which extends RL to decision-making involving multiple interacting agents (Busoniu et al., 2008; Hernandez-Leal et al., 2018; 2019; OroojlooyJadid & Hajinezhad, 2019; Zhang et al., 2021), We advocate MARL as a principled method for solving economical issues. In MARL, agents might have asymmetric roles such leader-follower structure (Bucarey et al., 2019; Bai et al., 2021; Zhong et al., 2021) which is related to our work, while previous works mainly focus on finding the Stackelberg-Nash equilibrium (Başar & Olsder, 1998). Among these works, our MARL application in economics is most related to Zhong et al. (2021) who also study a myopic follower setting. In contrast, in their work the followers aim to find Nash equilibrium while we hope to find competitive equilibrium in EE. Also, we study general function approximations which bears more generality when handling large state space.

Optimism and Online Reinforcement Learning. Our work is related to another flurry line of works studying online RL cooperated with optimism. wFor tabular setting where state space is finite, how to propose online RL algorithms achieving $\widetilde{\mathcal{O}}(\sqrt{K})$ online regret is thoroughly studied (Azar et al., 2017; Jin et al., 2018; Zhang et al., 2020). Adopting the principle of Optimism in the face of Uncertainty (OFU) (Auer et al., 2002; 2009; Jin et al., 2018; 2019), they overestimate action-value functions by adding a bonus to incentive exploration. When the state space is large or even continuous, the use of function approximation is necessary. Also based on OFU, there are several researches (Jin et al., 2019; Wang et al., 2019; Cai et al., 2020a) apply (generalized) linear function approximation on the transition kernel or action-value function and prove $\widetilde{\mathcal{O}}(\sqrt{K})$ online regret. Beyond linear setting, a recent line of works study RL with general function approximation (Ayoub et al., 2020; Cai et al., 2020b; Jin et al., 2021a). Based on the notion of eluder dimension introduced by Russo & Van Roy (2014) that characterizes the complexity of function class, Ayoub et al. (2020); Cai et al. (2020b) combine non-linear value target regression and OFU, proposing online RL algorithms with general function approximations achieving $\widetilde{\mathcal{O}}(\sqrt{K})$ online regret. Jin et al. (2021a) also achieve such a goal by proposing a more generalized complexity measure: Bellman eluder dimension. All works mentioned above study RL problem involving a single agent, which is different from our interested bilevel systems.

Pessimism and Offline Reinforcement Learning. Our works are also related to many literature concerning pessimism and offline RL in recent years (Liu et al., 2020; Rashidinejad et al., 2021; Jin et al., 2021b; Xie et al., 2021; Uehara & Sun, 2021). Different from online RL, the introduction of offline dataset leads to a potential distribution shift. When the dataset has no coverage guarantee, Buckman et al. (2020); Zanette (2021) find that the lower bound of offline RL could even be exponential. Rather than assuming a well-explored dataset in many previous literature (Antos et al., 2007; Munos & Szepesvári, 2008; Yang et al., 2020; Ross & Bagnell, 2012; Chen & Jiang, 2019), several works (Rajaraman et al., 2020; Kidambi et al., 2020; Jin et al., 2021b) adopt pessimism in model estimation and prove $\widetilde{\mathcal{O}}(K^{-1/2})$ suboptimality even under a partial coverage dataset. Liu et al. (2020) propose a pessimistic variant of fitted Q-learning algorithm (Antos et al., 2007), achieving the optimal policy within a restricted class of policies without assuming the dataset to be well-explored. Jin et al. (2021b) propose a provably efficient algorithm with the spirit of pessimism to solve offline RL with linear function approximations, under no coverage assumption on the dataset. Rashidinejad et al. (2021) study the offline RL in the tabular case through lower confidence bound (LCB), only assuming the partial coverage assumption on the dataset. With general function approximations on offline RL and partial coverage dataset, the suboptimality bound $\widetilde{\mathcal{O}}(K^{-1/2})$ is achieved by both model-based (Uehara & Sun, 2021) and model-free (Xie et al., 2021) algorithms. They both apply Berstein inequality to sharpen the convergence rate to $\widetilde{\mathcal{O}}(K^{-1/2})$. Besides, all works mentioned above analyze the optimization problem over a single agent, different from a bilevel system.

C. Omitted Algorithms and Proof Sketches

C.1. Online Setting

C.1.1. Omitted Algorithms

We present the omitted algorithm ME (Algorithm 3) for model estimation and OPL (Algorithm 4) for optimistic planning respectively.

Algorithm 3 Model Estimation (ME)

Input: Feasible utility set $\mathcal{U}$ and transition set $\mathcal{P}$. Historical data $\{(s_h^\tau, a_h^\tau, b_h^\tau, \{u_h^{\tau,(i)}\}_{i\in[N]})\}_{(\tau,h)\in[k-1]\times[H]}$ and value function estimators $\{V_{h+1}^{\tau,(i)}\}_{(\tau,h,i)\in[k-1]\times[H]\times[N]}$. Optimism parameter $\beta^{(1)}$, $\beta^{(2)}$.

1: **for** $h = 1$ to H **do**
2: $u_h^{k,(i)} = \operatorname{argmin}_{u\in\mathcal{U}} \sum_{\tau=1}^{k-1}(u_h^{\tau,(i)} - u(s_h^\tau, x_h^{\tau,(i)}))^2$, for all $i \in [N]$.
3: $\mathcal{U}_h^{k,(i)} = \{u \in \mathcal{U} : \sum_{\tau=1}^{k-1}(u_h^{k,(i)}(s_h^\tau, x_h^{\tau,(i)}) - u(s_h^\tau, x_h^{\tau,(i)}))^2 \le \beta^{(1)}\}$, for all $i \in [N]$.
4: $P_h^k = \operatorname{argmin}_{P\in\mathcal{P}} \sum_{\tau=1}^{k-1}(\sum_{i=1}^N V_{h+1}^{\tau,(i)}(s_{h+1}^\tau) - \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{\tau,(i)}(s')P(\mathrm{d}s'|s_h^\tau, b_h^\tau))^2$.
5: $\mathcal{P}_h^k = \{P \in \mathcal{P} : \sum_{\tau=1}^{k-1}(\int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{\tau,(i)}(s')P_h^k(s'|s_h^\tau, b_h^\tau)\mathrm{d}s' - \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{\tau,(i)}(s')P(s'|s_h^\tau, b_h^\tau))^2 \le \beta^{(2)}\}$.
6: **end for**
7: **Return** $\{\mathcal{U}_h^{k,(i)}\}_{(h,i)\in[H]\times[N]}$ and $\{\mathcal{P}_h^k\}_{h\in[H]}$.

Algorithm 4 Optimistic Planning (OPL)

Input: Utility confidence sets $\{\mathcal{U}_h^{k,(i)}\}_{(h,i)\in[H]\times[N]}$ and transition confidence sets $\{\mathcal{P}_h^k\}_{h\in[H]}$.

1: **for** $h = H$ to 1 **do**
2: $\widehat{u}_h^{k,(i)}(\cdot,\cdot) = \arg\max_{u\in\mathcal{U}_h^{k,(i)}} u(\cdot,\cdot)$, $\forall i \in [N]$.
3: $\widehat{P}_h^k(\cdot|\cdot,\cdot) = \arg\max_{P\in\mathcal{P}_h^k} \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{k,(i)}(s')P(\mathrm{d}s'|\cdot,\cdot)$.
4: $\nu_h^k(\cdot) = \mathtt{CE}(\{\widehat{u}_h^{k,(i)}(\cdot,\cdot)\}_{i\in[N]})$.
5: $\pi_h^k(\cdot) = \arg\max_{b\in\mathcal{B}} \sum_{i=1}^N \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s')\widehat{P}_h^k(\mathrm{d}s'|\cdot, b)$.
6: $Q_h^{k,(i)}(\cdot,\cdot,\cdot) = \widehat{u}_h^{k,(i)}(\cdot,\cdot) + \mathtt{Clip}_{[0,H-h]}\{\int_{\mathcal{S}} V_{h+1}^{k,(i)}(s')\widehat{P}_h^k(\mathrm{d}s'|\cdot,\cdot)\}$, $\forall i \in [N]$.
7: $V_h^{k,(i)}(\cdot) = Q_h^{k,(i)}(\cdot, \nu_h^{k,(i)}(\cdot), \pi_h^k(\cdot))$, $\forall i \in [N]$.
8: **end for**

C.1.2. Proof Sketch of Theorem 3.4

In the sequel, we sketch the proof of the first conclusion of Theorem 3.4, i.e., the upper bound on the regret for joint optimality. Missing details are left to Appendix F. The proof of the regret for fair division property is left to Appendix F.4. We start from a decomposition of the online learning regret in the following lemma.

Lemma C.1 (Regret Decomposition)**.** *We can decompose the online regret defined in (15) as following,*

$$\begin{aligned}\mathrm{Regret}(K) = &\sum_{k=1}^K\sum_{h=1}^H\sum_{i=1}^N \mathbb{E}_{\pi^\dagger(\nu^k),\nu^\star(\nu^k)}\left[Q_h^{k,(i)}(s_h^k, \nu_h^{\star,(i)}(\nu^k)(s_h^k), \pi_h^\dagger(\nu^k)(s_h^k)) - Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^k(s_h^k))\right]\\ &+ \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^\dagger(\nu^k),\nu^\star(\nu^k)}\left[\iota_h^k(s_h^k, a_h^k, b_h^k)\right] + \sum_{k=1}^K\sum_{i=1}^N V_1^{k,(i)}(s_1^k) - V_1^{(\pi^k,\nu^k),(i)}(s_1^k),\end{aligned} \tag{33}$$

where $\iota_h^k(\cdot,\cdot,\cdot)$ is defined as for any $(s_h, a_h, b_h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$,

$$\iota_h^k(s_h, a_h, b_h) = \sum_{i=1}^N u_h^{(i)}(s_h, x_h^{(i)}) + \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s')P_h(\mathrm{d}s'|s_h, b_h) - Q_h^{k,(i)}(s_h, x_h^{(i)}, b_h). \tag{34}$$

Here the functions $V_h^{k,(i)}$, $Q_h^{k,(i)}$, and the policies (π_h^k, ν_h^k) are selected by Algorithm 1.

Proof of Lemma C.1. See Appendix F.1 for a detailed proof. □

Therefore, it suffices to establish upper bounds for each term on the right-hand side of (33). The first term is characterized by the following lemma, which is derived from the choice of (π^k, ν^k) on each episode.

Lemma C.2 (One-Step Competitive Equilibrium and Social Welfare Maximization)**.** *According to Algorithm 1, for any $(k,h) \in [K] \times [H]$ and any $s_h \in \mathcal{S}$, it holds that*

$$\sum_{i=1}^{N} Q_h^{k,(i)}(s_h^k, \nu_h^{\star,(i)}(\nu^k)(s_h^k), \pi_h^{\dagger}(\nu^k)(s_h^k)) - Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^k(s_h^k)) \le 0. \tag{35}$$

Proof of Lemma C.2. See Appendix F.2 for a detailed proof. □

Besides, the second and the third terms of the right-hand side of (33) are characterized by the next lemma.

Lemma C.3 (Optimism and Accuracy)**.** *By setting the optimism parameter $\beta^{(1)}$ and $\beta^{(2)}$ as*

$$\beta^{(1)} = 2\log\left(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty) \cdot 2NH/\delta\right) + 4\big(1 + \sqrt{\log(8K^2H/\delta)}\big), \tag{36}$$

$$\beta^{(2)} = 2H^2N^2 \cdot \log\left(\mathcal{N}(1/(KHN), \mathcal{P}, \|\cdot\|_{\infty,1}) \cdot 2H/\delta\right) + 4\big(HN + \sqrt{H^2N^2/4 \cdot \log(8K^2H/\delta)}\big). \tag{37}$$

in Algorithm 1, then with probability at least $1-\delta$, the following two things holds.
(1) (Optimism) For all $(k,h) \in [K] \times [H]$ and any $(s_h, a_h, b_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, it holds that

$$\sum_{i=1}^{N} u_h^{(i)}(s_h, x_h^{(i)}) + \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s') P_h(\mathrm{d}s' | s_h, b_h) - Q_h^{k,(i)}(s_h, x_h^{(i)}, b_h) \le 0. \tag{38}$$

(2) (Accuracy) By denoting $d = \max\{\dim_{\mathbb{E}}(\mathcal{U}, 1/K), \dim_{\mathbb{E}}(\mathcal{Z}_{\mathcal{P}}, 1/K)\}$, it holds that

$$\sum_{k=1}^{K}\sum_{i=1}^{N} V_1^{k,(i)}(s_1^k) - V_1^{(\pi^k,\nu^k),(i)}(s_1^k) \le \mathcal{O}\big(\sqrt{KH^3N^2\log(4/\delta)}\big) + H\sqrt{d(N^2\beta^{(1)} + \beta^{(2)})K}\big) + dH^2N. \tag{39}$$

Proof of Lemma C.3. See Appendix F.3 for a detailed proof. □

Proof of Theorem 3.4. Combining Lemma C.1, Lemma C.2, and Lemma C.3, we can prove Theorem 3.4. According to Lemma C.1, Lemma C.2, and Lemma C.3, with probability at least $1-\delta$, it holds that

$$\begin{aligned}\mathrm{Regret}(K) &\le \sqrt{8KH^3N^2\log(4/\delta)} + 4H\sqrt{2d(N^2\beta^{(1)} + \beta^{(2)})K} + dH^2N \\ &\le \mathcal{O}\big(\sqrt{dH^2N^2(\beta^1 + \beta^2)K}\big),\end{aligned}$$

which finishes the proof of Theorem 3.4. □

C.2. Offline Setting

Our proof is based on the following two key lemmas.

Lemma C.4 (Upper Bound of Suboptimality)**.** *For the output $(\widehat{\pi}, \widehat{\nu})$ of Algorithm 2, it holds that,*

$$\begin{aligned}\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}) \le \sqrt{C_\rho^\star} \cdot \sum_{h=1}^{H}\left(\sqrt{\epsilon_h^{\widehat{u}}} + HN \cdot \sqrt{\epsilon_h^{\widehat{P}}}\right) \\ + \sum_{i=1}^{N}\left(\widehat{V}_{1,(\widehat{P},\widehat{u})}^{(\widehat{\pi},\widehat{\nu}),(i)} - V_1^{(\widehat{\pi},\widehat{\nu}),(i)}\right)(s_1),\end{aligned}$$

where error terms are defined as $\epsilon_h^{\widehat{u}} := \mathbb{E}_{\rho_h}\sum_{i=1}^{N}|\widehat{u}_h^{(i)}(s, x^{(i)}) - u_h(s, x^{(i)})|^2$ and $\epsilon_h^{\widehat{P}} := \mathbb{E}_{\rho_h}\|\widehat{P}_h(\cdot \,|\, s, b) - P_h(\cdot \,|\, s, b)\|_1^2$.

Proof of Lemma C.4. See Appendix G.1 for detailed proof. □

In the sequel, we bound the two summations in the suboptimality upper bound in Theorem C.4 respectively. To this end, we introduce the following results, concluded in Lemma C.5, Theorem C.6, and Theorem C.7.

Lemma C.5 (Pessimistic)**.** *Under event* $E := \{u_h^{(i)} \in \mathcal{U}_{h,\xi_1}^{(i)}, P_h \in \mathcal{P}_{h,\xi_2}$, *for all* $(i,h) \in [N] \times [H]\}$, *it holds that,*

$$\sum_{i=1}^{N} \widehat{V}_{1,(\widehat{P},\widehat{u})}^{(\widehat{\pi},\widehat{\nu}),(i)}(s_1) - V_1^{(\widehat{\pi},\widehat{\nu}),(i)}(s_1) \leq 0.$$

Proof of Lemma C.5. See Appendix G.2 for detailed proof. □

Based on Lemma C.4 and Lemma C.5, what remains is to upper bound $\epsilon_h^{\widehat{u}}$ and $\epsilon_h^{\widehat{P}}$ in Lemma C.4 respectively and to show that the event E holds with high probability. These are shown by the following two theorems.

Theorem C.6 (Analysis for Utility Function Estimation)**.** *For the output* $(\widehat{\pi},\widehat{\nu})$ *of Algorithm 2, the following statements hold with probability at least* $1-\delta/2$*,*

1. $E_0 := \{u_h^{(i)} \in \mathcal{U}_{h,\xi_1}^{(i)}$, *for all* $(i,h) \in [N] \times [H]\}$ *holds.*
2. $\epsilon_h^{\widehat{u}} \leq \mathcal{O}(\log(\mathcal{N}(1/K^2, \mathcal{U}, \|\cdot\|_\infty) \cdot NH/\delta)N/K)$*, for all* $(i,h) \in [N] \times [H]$.

Proof of Theorem C.6. See Appendix G.3 for detailed proof. □

Theorem C.7 (Analysis for Transition Kernel Estimation)**.** *For the output* $(\widehat{\pi},\widehat{\nu})$ *of Algorithm 2,the following statements hold with probability at least* $1-\delta/2$*,*

1. $E_1 := \{P_h \in \mathcal{P}_{h,\xi_2}$, *for all* $h \in [H]\}$ *holds.*
2. $\epsilon_h^{\widehat{P}} \leq \mathcal{O}(\log(\mathcal{N}_{[]}(1/K^2, \mathcal{P}, \|\cdot\|_{2,\infty})H/\delta)/K)$ *for all* $h \in [H]$.

Proof of Theorem C.7. See Appendix G.4 for detailed proof. □

Now combining the result of Theorem C.7 and Theorem C.6 and noting that $E = E_0 \cap E_1$, we can show that with probability at least $1-\delta$, the event E holds, which implies that the conclusion of Lemma C.5 holds. Meanwhile, error terms defined in Lemma C.4 are bounded as follows

$$\epsilon_h^{\widehat{u}} \leq \mathcal{O}(N\iota/K), \qquad \epsilon_h^{\widehat{P}} \leq \mathcal{O}(\iota/K),$$

where we define $\iota = \log \mathcal{N}_{[]}(1/K^2, \mathcal{P}, \|\cdot\|_{1,\infty}) + \log \mathcal{N}(1/K^2, \mathcal{U}, \|\cdot\|_\infty) + \log(HN/\delta)$.

Finally, according to Lemma C.4, we have that

$$\mathrm{SubOpt}(\widehat{\pi},\widehat{\nu},s_1) \leq \sqrt{C_\rho^\star} \cdot \sum_{h=1}^{H} \left(\sqrt{\epsilon_h^{\widehat{u}}} + HN \cdot \sqrt{\epsilon_h^{\widehat{P}}} \right) \leq \mathcal{O}(\sqrt{H^4N^2\iota C_\rho^\star/K}),$$

which finishes the proof of Theorem 4.4.

D. Special Cases

D.1. Linear Function Approximation

On the first case, we parameterize $\mathcal{P}$ and $\mathcal{U}$ by a common parameter vector $\theta \in \mathbb{R}^d$. We assume there exist an absolute constant d, known feature maps ϕ and $\{\psi_i\}_{i\in[N]}$, such that $\mathcal{P} = \{P(s' \mid s,b) = \phi(s',s,b)^\top\theta) : \theta \in \Theta\}$ and $\mathcal{U} = \{u(s,x^{(i)}) = \psi_i(s,x^{(i)})^\top\theta : \theta \in \Theta\}$. Following Russo & Van Roy (2013), we assert the following assumption.

Assumption D.1 (Regularity of Linear Function Approximation)**.** We assume the following two regularity conditions. (1) $\sup_{(s',s,b)\in\mathcal{S}\times\mathcal{S}\times\mathcal{B}} \|\phi(s',s,b)\|_2 \leq 1$ and $\sup_{(s,x^{(i)})\in\mathcal{S}\times\mathcal{X}^{(i)}} \|\psi_i(s,x^{(i)})\|_2 \leq 1$. (2) $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$.

Corollary D.2 (Theoretical Analysis of Algorithm 1 and Algorithm 2 with Linear Function Approximations)**.** *Under the same conditions as in Theorem 3.4, it holds with probability at least* $1-\delta$ *that the regret for joint optimality and for fair division property of Algorithm 1 satisfy,*

$$\mathrm{Regret}_{\mathrm{CE,SWM}}(K) \le \widetilde{\mathcal{O}}(\sqrt{d^2H^4N^4K}), \qquad \mathrm{Regret}_{\mathrm{FD}}(K) \le \widetilde{\mathcal{O}}(\sqrt{d^2H^2N^2K}).$$

For Algorithm 2, on the same condition as Theorem 4.4, it holds with probability at least $1-\delta$*,*

$$\mathrm{SubOpt}(\widehat{\pi},\widehat{\nu}) \le \widetilde{\mathcal{O}}(\sqrt{H^4N^2dC_\rho^\star/K}), \qquad \mathcal{L}_{\mathrm{FD}}(\widehat{\nu}) \le \widetilde{\mathcal{O}}(\sqrt{H^2N^2d/K}).$$

Proof of Corollary D.2. It suffices to upper bound the covering number, bracketing number, and eluder dimension under Assumption D.1 respectively. For the upper bound of covering number and bracketing number, we introduce the following lemma.

Lemma D.3 (Specification of Covering Number and Bracketing Number)**.** *Under Assumption D.1, it holds for all* $\epsilon \in (0,1)$ *that*

$$\log \mathcal{N}(\epsilon,\mathcal{U},\|\cdot\|_\infty) \le d\log(3h_{\mathrm{U}}/\epsilon) = \widetilde{\mathcal{O}}(d),$$
$$\log \mathcal{N}(\epsilon,\mathcal{P},\|\cdot\|_{1,\infty}) \le \log \mathcal{N}_{[]}(2\epsilon,\mathcal{P},\|\cdot\|_{1,\infty}) \le \log(4h_{\mathrm{U}}|\mathcal{S}|/\epsilon) = \widetilde{\mathcal{O}}(d).$$

Proof of Lemma D.3. For the first inequality in Lemma D.3, we prove it by definition of covering number. By the second point in Assumption D.1, Θ is a ball with radius 1 in d-dimension Euclidean space, which guarantees that (Lemma 5.2 of Vershynin (2010))

$$\log \mathcal{N}(\epsilon,\Theta,\|\cdot\|_\infty) \le d\log(3/\epsilon). \tag{40}$$

Taking the ϵ-covering of Θ as $\Theta_\epsilon = \{\theta_j\}_{j\in[M]}$ and arbitrarily fixing $i \in [N]$, it implies for any $u_\theta(\cdot,\cdot) = h(\psi_i(\cdot,\cdot)^\top\theta) \in \mathcal{U}$, there exists $j \in [M]$ such that

$$|u_\theta(s,x^{(i)}) - u_{\theta_j}(s,x^{(i)})| = |\psi_i(s,x^{(i)})^\top(\theta-\theta_j)| \le \|\psi_i(s,x^{(i)})\|_2\|\theta-\theta_j\|_2 \le \|\theta-\theta_j\|_2 \le \epsilon, \tag{41}$$

where the second last inequality relies on the first point in Assumption D.1 and the last inequality follows from the definition of covering number. Taking supreme on the both side of (41) over $(s,x^{(i)}) \in \mathcal{S}\times\mathcal{X}^{(i)}$, we show that $\{u_{\theta_j} = \phi(\cdot,\cdot)^\top\theta_j\}_{j\in[M]}$ is also a ϵ-covering of $\mathcal{U}$ under ∞-norm, implying that

$$\log \mathcal{N}(\epsilon',\mathcal{U}.\|\cdot\|_\infty) \le d\log(3/\epsilon') = \widetilde{\mathcal{O}}(d).$$

Hence we complete the proof of the first part of Lemma D.3.

As for the second part, we introduce the following key lemma to connect bracketing number with covering number, which is proved in Sen (2018).

Lemma D.4 (Theorem 2.14 in Sen (2018))**.** *Let* $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ *be a class of functions satisfying the following condition*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le d(\theta_1,\theta_2)\,F(x), \quad \forall x \in \mathcal{X}, \forall \theta_1,\theta_2 \in \Theta,$$

for some fixed function F *and metric* d*. Then, for any norm* $\|\cdot\|$*, it yields that*

$$\mathcal{N}_{[]}(2\epsilon\|F\|,\mathcal{F},\|\cdot\|) \le \mathcal{N}(\epsilon,\Theta,d).$$

Under Assumption D.1, for any two kernels $P_{\theta_1}, P_{\theta_2} \in \mathcal{P}$, it holds that

$$|P_{\theta_1}(s' \mid s,b) - P_{\theta_2}(s' \mid s,b)| = |\phi(s',s,b)^\top(\theta_1-\theta_2)| \le h_{\mathrm{U}}\|\phi(s',s,b)\|_2\|\theta_1-\theta_2\|_2$$

Applying Lemma D.4 with $F(s',s,b) = \|\phi(s',s,b)\|_2$ and noticing that $\|F\|_{1,\infty} \le |\mathcal{S}|$, we derive that

$$\log \mathcal{N}_{[]}(2|\mathcal{S}|\epsilon,\mathcal{P},\|\cdot\|_{1,\infty}) \le \log \mathcal{N}(\epsilon,\Theta,\|\cdot\|_2) \le d\log(1/\epsilon) = \widetilde{\mathcal{O}}(d),$$

where the last inequality follows from (40). Taking $\epsilon' = 2|\mathcal{S}|\epsilon$, we have that

$$\log \mathcal{N}_{[]}(\epsilon',\mathcal{P},\|\cdot\|_{1,\infty}) \le d\log(2|\mathcal{S}|/\epsilon').$$

Noting that for all normed space $(\mathcal{X},\|\cdot\|)$, it holds that $\mathcal{N}(\epsilon,\mathcal{X},\|\cdot\|) \le \mathcal{N}_{[]}(2\epsilon,\mathcal{X},\|\cdot\|)$, which concludes the proof of Lemma D.3. □

As the second step in the proof of Corollary D.2, under Assumption D.1 we upper bound the eluder dimension defined in 3.3 by the following lemma.

Lemma D.5 (Specification of Eluder Dimension)**.** *Under Assumption D.1, it holds for all $\epsilon \in (0,1)$ that*

$$\dim_{\mathbb{E}}(\mathcal{U}, \epsilon) \leq \widetilde{\mathcal{O}}(d), \qquad \dim_{\mathbb{E}}(\mathcal{Z}_{\mathcal{P}}, \epsilon) \leq \widetilde{\mathcal{O}}(d),$$

where $\mathcal{Z}_{\mathcal{P}}$ is defined in Section 3.3.

Proof of Lemma D.5. The proof is a special case of the following lemma proved in Russo & Van Roy (2013).

Lemma D.6 (Proposition 12 of Russo & Van Roy (2013))**.** *We define the function class on $\mathcal{X}$ as*

$$\mathcal{F} = \{h(\varphi(\cdot)^\top \theta) : \theta \in \mathbb{R}^d\} \subset \{f : \mathcal{X} \mapsto \mathbb{R}\},$$

for a fixed differential function $h(\cdot)$ and a feature map $\varphi(\cdot)$. If we assume that $0 < h_{\mathrm{L}} \leq h'(y) \leq h_{\mathrm{U}}$ for all $y \in \mathbb{R}$, then it holds for all $\epsilon \in (0,1)$ that

$$\dim_{\mathbb{E}}(\mathcal{F}, \epsilon) \leq \mathcal{O}\left(dr^2 \log\left(r^2 + \frac{r^2 h_{\mathrm{U}}^2 \cdot \sup_{\theta\in\Theta} \|\theta\|_2 \cdot \sup_{x\in\mathcal{X}} \|\varphi(x)\|_2}{\epsilon^2}\right)\right),$$

where $r := h_{\mathrm{U}}/h_{\mathrm{L}}$ and $\dim_{\mathbb{E}}(\mathcal{F}, \epsilon)$ is defined in Definition 3.3.

Now we are ready to specify $\dim_{\mathbb{E}}(\mathcal{U}, \epsilon)$ and $\dim_{\mathbb{E}}(\mathcal{Z}_P, \epsilon)$ by taking $h(\cdot)$ as identity function. Under Assumption D.1, it holds that $\|\theta\|_2 \leq 1$ for all $\theta \in \Theta$ and $\|\psi_i(\cdot,\cdot)\|_2 \leq 1$, which implies that

$$\dim_{\mathbb{E}}(\mathcal{U}, \epsilon) \leq \mathcal{O}\left(d \log\left(1 + \frac{1}{\epsilon^2}\right)\right) = \widetilde{\mathcal{O}}(d). \tag{42}$$

The analysis for $\dim_{\mathbb{E}}(\epsilon, \mathcal{Z}_{\mathcal{P}})$ needs more elaborations. Recall that for each $z_P \in \mathcal{Z}_{\mathcal{P}}$, it holds that

$$z_P((s,b,f)) = \int_{\mathcal{S}} f(s') P(\mathrm{d}s' \,|\, s, b) = \int_{\mathcal{S}} f(s') \phi(s', s, b)^\top \theta \mathrm{d}s' = \theta^\top \int_{\mathcal{S}} f(s') \phi(s', s, b) \mathrm{d}s',$$

where f is a arbitrary function in $f : \mathcal{S} \mapsto [0, HN]$ and $(s,b) \in \mathcal{S} \times \mathcal{B}$. If we take $\mathcal{X} = \mathcal{S} \times \mathcal{B} \times \{f : \mathcal{S} \mapsto [0, HN]\}$ and $\varphi((s,b,f)) = \int_{\mathcal{S}} f(s')\phi(s', s, b)\mathrm{d}s'$ in Lemma D.6, we obtain that

$$\dim_{\mathbb{E}}(\mathcal{Z}_{\mathcal{P}}, \epsilon) \leq \mathcal{O}\left(d \log\left(1 + HN \sup_{(s,b)\in\mathcal{S}\times\mathcal{B}} \int_{\mathcal{S}} \frac{|\phi(s', s, b)|}{\epsilon^2} \mathrm{d}s'\right)\right) \leq \mathcal{O}(d \log(1 + HN|\mathcal{S}|)) = \widetilde{\mathcal{O}}(d),$$

where the last inequality relies on the first point of Assumption D.1. □

Then we are ready to prove corollary D.2. Based on Lemma D.3, we can upper bound the optimism parameters (36) in Algorithm 1 as

$$\beta^{(1)} = \widetilde{\mathcal{O}}(d), \qquad \beta^{(2)} = \widetilde{\mathcal{O}}(H^2 N^2 d). \tag{43}$$

We also upper bound the pessimism parameters defined in Theorem 4.4 of Algorithm 2 as

$$\xi_1 = \widetilde{\mathcal{O}}(d/K), \qquad \xi_2 = \widetilde{\mathcal{O}}(d/K). \tag{44}$$

Combining Lemma D.5 and plugging them into Theorem 3.4, Theorem 4.4 respectively, we prove Corollary D.2. □

Remark D.7 (Generalized Linear Kernel Case). We remark that based on Lemma D.6, our conclusion can also be extended to the setting when utility functions are generalized linear (Guo et al., 2021).

Remark D.8 (Tabular Case). Let feature maps being the canonical basis on the factorized space, that is, $\phi(s', s, b) = \mathbf{e}_{(s',s,b)}$, $\psi_i(s, x^{(i)}) = \mathbf{e}_{(s,x^{(i)})}$, and $d = \max\{|\mathcal{S}|^2|\mathcal{B}|, \max_i |\mathcal{S}||\mathcal{X}^{(i)}|\}$, then Assumption D.1 is satisfied.

D.2. Reproducing Kernel Hilbert Space

We consider the case when utility functions $u_h^{(i)}$ and transition kernel P_h are parameterized by a subset of a reproducing kernel Hilbert space (RKHS). Specifically, we consider two RKHS's $\mathcal{H}^u$ and $\mathcal{H}^P$ associated with two positive definite kernels $\mathcal{K}^u : (\mathcal{S} \times \mathcal{X}^{(i)}) \times (\mathcal{S} \times \mathcal{X}^{(i)}) \mapsto \mathbb{R}_+$ and $\mathcal{K}^P : (\mathcal{S} \times \mathcal{B} \times \mathcal{S}) \times (\mathcal{S} \times \mathcal{B} \times \mathcal{S}) \mapsto \mathbb{R}_+$ respectively. We denote the corresponding feature mappings by $\phi^u : \mathcal{S} \times \mathcal{A}^{(1)} \mapsto \mathcal{H}^u$ and $\phi^P : \mathcal{S} \times \mathcal{B} \times \mathcal{S} \mapsto \mathcal{H}$. We assume that

$$\mathcal{U} = \big\{\langle \phi^u(\cdot,\cdot), f\rangle_{\mathcal{H}^u} : f \in \mathcal{H}^u, \|f\|_{\mathcal{H}^u} \le R^u\big\} = \big\{\langle \phi^u(\cdot,\cdot), f\rangle_{\mathcal{H}^u} : f \in \mathcal{H}^u_{R^u}\big\},$$

$$\mathcal{P} = \big\{\langle \phi^P(\cdot,\cdot,\cdot), f\rangle_{\mathcal{H}^P} : f \in \mathcal{H}^P, \|f\|_{\mathcal{H}^P} \le R^P\big\} = \big\{\langle \phi^P(\cdot,\cdot,\cdot), f\rangle_{\mathcal{H}^P} : f \in \mathcal{H}^P_{R^P}\big\}.$$

By Mercer's theorem (Steinwart & Christmann, 2008), we denote the decomposition of $\mathcal{K}^u$ and $\mathcal{K}^P$ as

$$\mathcal{K}(x,y) = \sum_{j=1}^{+\infty} \lambda_j^u \phi_j^u(x)\phi_j^u(y), \quad \mathcal{K}(x,y) = \sum_{j=1}^{+\infty} \lambda_j^P \phi_j^P(x)\phi_j^P(y),$$

where $x, y \in \mathcal{Y}$ with $\mathcal{Y} = \mathcal{S} \times \mathcal{X}^{(i)}$ for $\mathcal{U}$ and $\mathcal{Y} = \mathcal{S} \times \mathcal{B} \times \mathcal{S}$ for $\mathcal{P}$. Following Cai et al. (2020b), we assume that both $\mathcal{H}^u$ and $\mathcal{H}^P$ satisfy the following regularity conditions. For simplicity, we omit the superscript u or P.

Assumption D.9 (Regularity of RKHS)**.** We assume $\mathcal{K}$ satisfies the following two regularity conditions.
(1) It holds that $|\mathcal{K}(x,y)| \le 1, |\phi_j(x)| \le 1$, and $\lambda_j \le 1$ for any $x, y \in \mathcal{Y}$ and $j \in \mathbb{N}$.
(2) There exist a threshold $\gamma \in (0, 1/2)$ and constant $C_1, C_2 > 0$ such that $\lambda_j \le C_1 \cdot \exp(-C_2 j^\gamma)$ for any $j \in \mathbb{N}$.

Corollary D.10 (Theoretical Analysis of Algorithm 1 and Algorithm 2 with Kernel Function Approximations)**.** *Under the same conditions as in Theorem 3.4, it holds with probability at least $1-\delta$ that the regret for joint optimality and for fair division property of Algorithm 1 satisfy*

$$\mathrm{Regret}_{\mathrm{CE,SWM}}(K) \lesssim \mathcal{O}\big(H^2 N K^{1/2} \log^2(1/\gamma)/\gamma \cdot \log^{1+1/\gamma}(2|\mathcal{S}|RHNK/\delta)\big),$$

$$\mathrm{Regret}_{\mathrm{FD}}(K) \lesssim \mathcal{O}\big(HNK^{1/2} \log^2(1/\gamma)/\gamma \cdot \log^{1+1/\gamma}(RHNK/\delta)\big).$$

Besides, under the same conditions as in Theorem 4.4, it holds with probability at least $1-\delta$ that the suboptimality and offline FD loss of Algorithm 2 satisfy

$$\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}, s_1) \lesssim \mathcal{O}\big((C_\rho^\star)^{1/2} H^2 N \cdot K^{-1/2} \log(1/\gamma)/\gamma \cdot \log^{1/2+1/2\gamma}(2|\mathcal{S}|RHNK^2/\delta)\big),$$
$$\mathcal{L}_{FD}(\widehat{\nu}) \lesssim \mathcal{O}\big(HN \cdot K^{-1/2} \log(1/\gamma)/\gamma \cdot \log^{1/2+1/2\gamma}(RHNK^2/\delta)\big).$$

Proof of Corollary D.10. When $\mathcal{U}$ and $\mathcal{P}$ are both parameterized by RKHS, the covering numbers and bracketing numbers of $\mathcal{U}$ and $\mathcal{P}$, together with the eluder dimension d, can be upper bounded explicitly, which are concluded in the following two lemmas.

Lemma D.11 (Covering Number and Bracketing Number with RKHS)**.** *Under Assumption D.9, it holds for all $\epsilon \in (0,1)$ that*

$$\log \mathcal{N}(\epsilon, \mathcal{U}, \|\cdot\|_\infty) \le C_3 \cdot \log^2(1/\gamma)/\gamma \cdot \log^{1+1/\gamma}(R/\epsilon),$$
$$\log \mathcal{N}(\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty}) \le \log \mathcal{N}_{[]}(2\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty}) \le C_4 \cdot \log^2(1/\gamma)/\gamma \cdot \log^{1+1/\gamma}(2|\mathcal{S}|R/\epsilon).$$

where $C_3, C_4 > 0$ are absolute constants.

Proof of Lemma D.11. For notational simplicity, we omit all the superscripts u or P without making confusion. For function class $\mathcal{U}$, we invoke Lemma I.1 (Cai et al., 2020b) which shows that

$$\log \mathcal{N}(\epsilon, \mathcal{U}, \|\cdot\|_\infty) \le C_3 \cdot \log^2(1/\gamma)/\gamma \cdot \log^{1+1/\gamma}(R/\epsilon),$$

for some absolute constant $C_3 > 0$. In the sequel, we deal with function class $\mathcal{P}$ and we start from bounding the bracketing number $\mathcal{N}_{[]}(2\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty})$ since it upper bounds the covering number $\mathcal{N}(\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty})$. We first note that

$\mathcal{N}_{[]}(2\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty}) \le \mathcal{N}_{[]}(2\epsilon/|\mathcal{S}|, \mathcal{P}, \|\cdot\|_\infty)$. Now we apply a truncation argument. Let d_0 be an integer which will be specified later. Denote $\widetilde{\mathcal{P}}$ as

$$\widetilde{\mathcal{P}} = \left\{ P = \sum_{j=1}^{d_0} v_j \sqrt{\lambda_j}\phi_j : \|v\|_2 \le R \right\} \subseteq \mathcal{P},$$

which is in fact a linear function class over $\mathcal{Y} = \mathcal{S}\times\mathcal{B}\times\mathcal{S}$ with finite dimension $d_0 < \infty$. Also, for any $P \in \mathcal{P}$, we define the truncation of P to finite dimensional space $\widetilde{\mathcal{P}}$ as

$$\widetilde{P} = \sum_{j=1}^{d_0} \langle P, \sqrt{\lambda_j}\phi_j\rangle_{\mathcal{H}} \sqrt{\lambda_j}\phi_j \in \widetilde{\mathcal{P}}.$$

The difference between P and $\widetilde{P}$ under the $\|\cdot\|_\infty$-norm can be bounded as

$$\begin{aligned}\|P - \widetilde{P}\|_\infty &= \sup_{y\in\mathcal{Y}} \left| \sum_{j=d_0+1}^{+\infty} \langle P, \sqrt{\lambda_j}\phi_j\rangle_{\mathcal{H}} \sqrt{\lambda_j}\phi_j(y) \right| \\ &\le \sum_{j=d_0+1}^{+\infty} \sqrt{\lambda_j}\|P\|_{\mathcal{H}} \|\sqrt{\lambda_j}\phi_j\|_{\mathcal{H}} \sup_{y\in\mathcal{Y}} |\phi_j(y)| \le R\cdot \sum_{j=d_0+1}^{+\infty} \sqrt{\lambda_j},\end{aligned}$$

from which we denote $\varepsilon_{d_0} = R\cdot\sum_{j=d_0+1}^{+\infty}\sqrt{\lambda_j}$ which is bounded later. Now let $\mathcal{S}_{d_0} = \{[\widetilde{g}_j^L, \widetilde{g}_j^U]\}_{j\in[\mathcal{N}_{[]}(\epsilon/|\mathcal{S}|, \widetilde{\mathcal{P}}, \|\cdot\|_\infty)]}$ be a smallest $\epsilon/|\mathcal{S}|$-bracket cover of $\widetilde{\mathcal{P}}$ under $\|\cdot\|_\infty$ norm. By the definition of bracketing in Section A, for any $P\in\mathcal{P}$, there exists a bracket $[\widetilde{g}_j^L, \widetilde{g}_j^U] \in \mathcal{S}_{d_0}$ such that $\widetilde{l}(y) \le \widetilde{P}(y) \le \widetilde{u}(y)$ for any $y\in\mathcal{Y}$. As a result,

$$\widetilde{g}_j^L - \varepsilon_{d_0} \le P(y) \le \widetilde{g}_j^U + \varepsilon_{d_0}, \quad \forall y\in\mathcal{Y}.$$

Define functions $g_j^L = \widetilde{g}_j^L - \varepsilon_{d_0}$ and $g_j^U = \widetilde{g}_j^U + \varepsilon_{d_0}$ respectively, and let $\mathcal{S}$ be the collect of the brackets $[g_j^L, g_j^U]$. Then the set $\mathcal{S}$ is an $(\epsilon/|\mathcal{S}| + 2\varepsilon_{d_0})$-bracket cover of $\mathcal{P}$ with $|\mathcal{S}| = |\mathcal{S}_{d_0}| = \mathcal{N}_{[]}(\epsilon/|\mathcal{S}|, \widetilde{\mathcal{P}}, \|\cdot\|_\infty)$. Thus we have that

$$\mathcal{N}_{[]}(\epsilon/|\mathcal{S}| + 2\varepsilon_{d_0}, \mathcal{P}, \|\cdot\|_{1,\infty}) \le |\mathcal{S}| = \mathcal{N}_{[]}(\epsilon/|\mathcal{S}|, \widetilde{\mathcal{P}}, \|\cdot\|_\infty).$$

By Lemma D.3, we know that $\log\mathcal{N}_{[]}(\epsilon/|\mathcal{S}|, \widetilde{\mathcal{P}}, \|\cdot\|_\infty) \le d_0\log(4d_0|\mathcal{S}|/\epsilon)$. Consequently, it then suffices to chooce a proper integar d_0 such that $2\varepsilon_{d_0} \le \epsilon/|\mathcal{S}|$. According to Lemma I.3 (Cai et al., 2020b), by choosing

$$d_0 = \lceil \widetilde{C}\cdot\log(1/\gamma)/\gamma\cdot\log^{1/\gamma}(2|\mathcal{S}R/\epsilon)\rceil,$$

where γ is specified in Assumption D.9, it holds that $\varepsilon_{d_0} \le \epsilon/2|\mathcal{S}|$. Therefore, we conclude that

$$\begin{aligned}\mathcal{N}_{[]}(2\epsilon/|\mathcal{S}|, \mathcal{P}, \|\cdot\|_{1,\infty}) &\le d_0\log(4d_0|\mathcal{S}|/\epsilon) \\ &\le C_4\cdot\log^2(1/\gamma)/\gamma\cdot\log^{1+1/\gamma}(2|\mathcal{S}|R/\epsilon).\end{aligned}$$

Finally, due to the fact that $\mathcal{N}(\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty}) \le \mathcal{N}_{[]}(2\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty})$, we can finish the proof of Lemma D.11. □

Lemma D.12 (Eluder Dimemsion with RKHS)**.** *Under Assumption D.9, there exists an absolute constant $C_5 > 0$ such that*

$$d = \max\{\dim_{\mathbb{E}}(\mathcal{U}, 1/K), \dim_{\mathbb{E}}(\mathcal{Z}_{\mathcal{P}}, 1/K)\} \le C_5\cdot\log^2(1/\gamma)/\gamma\cdot\log^{1+1/\gamma}(RHNK).$$

Proof of Lemma D.12. See Lemma C.1 in Cai et al. (2020b) for a detailed proof. □

Combining Lemma D.11, D.12 with Theorem 3.4 and 4.4 finishes the proof of Corollary D.10. □

E. Proofs for Competitive Equilibrium and Social Welfare Maximization

E.1. Proof for Theorem 2.4

Proof of Theorem 2.4. We first show the inequality by induction. For $h = H$, it holds that for any $s_H \in \mathcal{S}$,

$$\begin{aligned} V_H^{(\pi,\nu^\star(\nu)),(i)}(s_H) &= Q_H^{(\pi,\nu^\star(\nu)),(i)}(s_H, \nu_H^{\star,(i)}(\nu)(s_H), \pi_H(s_H)) \\ &= \max_{x_H^{(i)} \in \mathcal{X}^{(i)} : (\nu_H^p(s_H))^\top x_H^{(i)} \le (\nu_H^p(s_H))^\top e_H} u_H^{(i)}(s_H, x_H^{(i)}) \\ &\ge u_H^{(i)}(s_H, \nu_H^{(i)}(s_H)) = V_H^{(\pi,\nu),(i)}(s_H). \end{aligned} \tag{45}$$

This shows step H. Suppose that the inequality holds for step $h+1$, i.e., $V_{h+1}^{(\pi,\nu^\star(\nu)),(i)}(s_{h+1}) \ge V_{h+1}^{(\pi,\nu),(i)}(s_{h+1})$ for any $s_{h+1} \in \mathcal{S}$. Then for step h, we first have that for any $(s_h, x_h^{(i)}, b_h) \in \mathcal{S} \times \mathcal{X}^{(i)} \times \mathcal{B}$,

$$\begin{aligned} Q_h^{(\pi,\nu^\star(\nu)),(i)}(s_h, x_h^{(i)}, b_h) &= u_h^{(i)}(s_h, x_h^{(i)}) + \int_{\mathcal{S}} V_{h+1}^{(\pi,\nu^\star(\nu)),(i)}(s') P_h(\mathrm{d}s' | s_h, b_h) \\ &\ge u_h^{(i)}(s_h, x_h^{(i)}) + \int_{\mathcal{S}} V_{h+1}^{(\pi,\nu),(i)}(s') P_h(\mathrm{d}s' | s_h, b_h) \\ &= Q^{(\pi,\nu),(i)}(s_h, x_h^{(i)}, b_h), \end{aligned} \tag{46}$$

where the inequality follows by induction. Then we have that for any $s_h \in \mathcal{S}$,

$$\begin{aligned} V_h^{(\pi,\nu^\star(\nu)),(i)}(s_h) &= Q_h^{(\pi,\nu^\star(\nu)),(i)}(s_h, \nu_h^{\star,(i)}(\nu)(s_h), \pi_h(s_h)) \\ &\ge Q^{(\pi,\nu),(i)}(s_h, \nu_h^{\star,(i)}(\nu)(s_h), \pi_h(s_h)) \\ &= u_h^{(i)}(c_h, \nu_h^{\star,(i)}(\nu)(s_h)) + \int_{\mathcal{S}} V_{h+1}^{(\pi,\nu),(i)}(s') P_h(\mathrm{d}s' | s_h, \pi_h(s_h)) \\ &= \max_{x_h^{(i)} \in \mathcal{X}^{(i)} : (\nu_h^p(s_h))^\top x_h^{(i)} \le (\nu_h^p(s_h))^\top e_H} u_h^{(i)}(s_h, x_h^{(i)}) + \int_{\mathcal{S}} V_{h+1}^{(\pi,\nu),(i)}(s') P_h(\mathrm{d}s' | s_h, \pi_h(s_h)) \\ &\ge u_h^{(i)}(s_h, \nu_h^{(i)}(s_h)) + \int_{\mathcal{S}} V_{h+1}^{(\pi,\nu),(i)}(s') P_h(\mathrm{d}s' | s_h, \pi_h(s_h)) \\ &= V_h^{(\pi,\nu),(i)}(s_h), \end{aligned} \tag{47}$$

where the first inequality follows from (46) and the second inequality follows from the definition of $\nu_h^{\star,(i)}(\nu)$ and the fact that both ν and $\nu^\star(\nu)$ satisfy the resource constraints. This proves the first conclusion of Theorem 2.4. When the inequality holds for $h = 1$, from the previous proofs we know that all the inequalities become equalities, which further implies that $\nu_h(s_h)$ is a competitive equilibrium with respect to $\{u_h^{(i)}(s_h, \cdot)\}_{i \in [N]}$. □

E.2. Proof for Theorem 2.6

Proof of Theorem 2.6. We show the inequality by induction. For $h = H$, it holds that for any $s_h \in \mathcal{S}$,

$$\sum_{i=1}^N V_H^{(\pi^\star(\nu),\nu),(i)}(s_H) = \sum_{i=1}^N u_H^{(i)}(s_H, \nu_H^{(i)}(s_H)) = \sum_{i=1}^N V_H^{(\pi,\nu),(i)}(s_H). \tag{48}$$

Now we suppose that the inequality holds for step $h+1$, i.e., $\sum_{i=1}^N V_h^{(\pi^\star(\nu),\nu),(i)}(s_{h+1}) \geq \sum_{i=1}^N V_h^{(\pi,\nu),(i)}(s_{h+1})$ for any $s_{h+1} \in \mathcal{S}$. Then for step h we have that for any $s_h \in \mathcal{S}$,

$$\begin{aligned}
\sum_{i=1}^N V_h^{(\pi^\star(\nu),\nu),(i)}(s_h) &= \sum_{i=1}^N Q_h^{(\pi^\star(\nu),\nu),(i)}(s_h, \nu_h^{(i)}(s_h), \pi_h^\star(\nu)(s_h)) \\
&= \sum_{i=1}^N u_h^{(i)}(s_h, \nu_h^{(i)}(s_h)) + \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{(\pi^\star(\nu),\nu),(i)}(s') P_h(\mathrm{d}s'|s_h, \pi_h^\star(\nu)(s_h)) \\
&\geq \sum_{i=1}^N u_h^{(i)}(s_h, \nu_h^{(i)}(s_h)) + \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{(\pi^\star(\nu),\nu),(i)}(s') P_h(\mathrm{d}s'|s_h, \pi_h(s_h)) \\
&\geq \sum_{i=1}^N u_h^{(i)}(s_h, \nu_h^{(i)}(s_h)) + \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{(\pi,\nu),(i)}(s') P_h(\mathrm{d}s'|s_h, \pi_h(s_h)) \\
&= \sum_{i=1}^N V_h^{(\pi,\nu),(i)}(s_h).
\end{aligned} \tag{49}$$

where the first inequality follows from the choice of $\pi_h^\star$ in (7) and the second inequality follows from induction, proving the first part of Theorem 2.6. When the inequality holds for $h=1$, all the inequalities become equalities, which further implies that $\pi_h(s_h) \in \arg\max_{b_h \in \mathcal{B}} \sum_{i=1}^N V_{h+1}^{(\pi^\star(\nu),\nu),(i)}(s') P_h(\mathrm{d}s'|s_h, b_h)$, finishing the proof. □

F. Proofs for Online Learning Algorithm: Section 3

F.1. Proof for Lemma C.1

Proof of Lemma C.1. See Lemma 4.9 in Cai et al. (2020b) for a detailed proof. □

F.2. Proof for Lemma C.2

Proof of Lemma C.2. We decompose the left-hand side of (35) into to two terms

$$\begin{aligned}
&\sum_{i=1}^N Q_h^{k,(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h^k), \pi_h^\dagger(\nu^k)(s_h^k)) - Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^k(s_h^k)) \\
&= \underbrace{\sum_{i=1}^N Q_h^{k,(i)}(s_h^k, \nu_h^{\star,(i)}(\nu^k)(s_h^k), \pi_h^\dagger(\nu^k)(s_h^k)) - Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^\dagger(\nu^k)(s_h^k))}_{\text{(i)}} \\
&\quad + \underbrace{\sum_{i=1}^N Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^\dagger(\nu^k)(s_h^k)) - Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^k(s_h^k))}_{\text{(ii)}}.
\end{aligned} \tag{50}$$

For term (i), consider that for any agent $i \in [N]$, it holds that

$$\begin{aligned}
\text{(i)} &= Q_h^{k,(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h), \pi_h^\dagger(\nu^k)(s_h)) - Q_h^{k,(i)}(s_h, \nu_h^{k,(i)}(s_h), \pi_h^\star(\nu^k)(s_h)) \\
&= \widehat{u}_h^{k,(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h)) - \widehat{u}_h^{k,(i)}(s_h, \nu_h^{k,(i)}(\nu^k)(s_h)) \leq 0,
\end{aligned} \tag{51}$$

where the inequality holds due to the fact that ν_h^k is a competitive equilibrium policy against $\{\widehat{u}_h^{k,(i)}\}_{i\in[N]}$ and the definition of the best responce policy $\nu^\star(\nu^k)$ for the agent ν^k. For term (ii), consider we have that

$$\begin{aligned}
\text{(ii)} &= \sum_{i=1}^N Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^\dagger(\nu^k)(s_h^k)) - Q_h^{k,(i)}(s_h^k, \nu_h^{k,(i)}(s_h^k), \pi_h^k(s_h^k)) \\
&= \sum_{i=1}^N \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s') \widehat{P}_h^k(\mathrm{d}s'|s_h, \pi_h^\dagger(\nu^k)(s_h)) - \sum_{i=1}^N \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s') \widehat{P}_h^k(\mathrm{d}s'|s_h, \pi_h^k(s_h)) \leq 0,
\end{aligned} \tag{52}$$

where the inequality holds due to the greedy choice of π_h^k in Line 4 of Algorithm 1. This finished the proof. □

F.3. Proof for Lemma C.3

Proof of Lemma C.3. First we introduce the following definition of filtration for the later analysis.

Definition F.1 (Filtration: Online Learning)**.** We define the time index map $t(\cdot,\cdot)$ by $t(k,h) = H\cdot(j-1)+h$ for any $(k,h)\in[K]\times[H]$, which is a bijection from $[K]\times[H]$ to $[KH]$. Then, for any $(k,h)\in[K]\times[H]$, we define $\mathcal{F}_{t(k,h)}$ as the σ-algebra generated by

$$\left(s_1^1, a_1^1, b_1^1, \{u_h^{1,(i)}\}_{i\in[N]}, \cdots, s_H^1, a_H^1, b_H^1, \{u_H^{1,(i)}\}_{i\in[N]}, s_1^2, a_1^2, b_1^2, \{u_h^{2,(i)}\}_{i\in[N]}, \cdots, s_h^k, a_h^k, b_h^k, \{u_h^{k,(i)}\}_{i\in[N]}\right),$$

which are the utility samples and state-action pairs determined before s_{h+1}^k. The sequence $\{\mathcal{F}_t\}_{t\geq 1}$ is a filtration.

Also, we define E as the event when the true model is contained in the confidence set of Algorithm 1.

$$E := \left\{P_h \in \mathcal{P}_h^k, u_h^{(i)} \in \mathcal{U}_h^{k,(i)}, \quad \text{for all } (k,h,i)\in[K]\times[H]\times[N]\right\}. \tag{53}$$

Then we introduce the following lemma to show that event E happens with at least $1-p$ probability.

Lemma F.2. *For any $\delta\in[0,1]$, if we set*

$$\begin{aligned}\beta^{(1)} =& 2\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)\cdot 4HN/\delta) + 4(1+\sqrt{\log(16K^2HN/\delta)}),\\ \beta^{(2)} =& 2H^2N^2\log(\mathcal{N}(1/(KHN),\mathcal{P},\|\cdot\|_\infty)\cdot 4H/\delta) + 4HN(1+\sqrt{\log(16K^2H/\delta)}/2))\end{aligned}$$

in Algorithm 1, then with probability at least $1-\delta/2$, event E happens.

Proof of Lemma F.2. Let $\{(X_\tau, Y_\tau)\}_{\tau\geq 1}$ be a sequence of random elements in $\mathcal{X}\times\mathbb{R}$ for some measurable set $\mathcal{X}$. Let $\mathcal{Z}$ be a set of $[0,C]$-valued measurable functions with domain $\mathcal{X}$ for some constant $C>0$. Let $\mathcal{F} = \{\mathcal{F}_\tau\}_{\tau\geq 1}$ be a filtration such that for all $\tau\geq 1, (X_1, Y_1, \cdots, X_{\tau-1}, Y_{\tau-1}, X_\tau)$ is $\mathcal{F}_{\tau-1}$-measurable and there exists $z^*\in\mathcal{Z}$ such that $\mathbb{E}[Y_\tau|\mathcal{F}_{\tau-1}] = z^*(X_\tau)$ holds. The least-squares predictor given $\{(X_\tau,Y_\tau)\}_{\tau=1}^t$ is defined as

$$\widehat{z}_t = \operatorname*{argmin}_{z\in\mathcal{Z}} \sum_{\tau=1}^t (z(X_\tau) - Y_\tau)^2.$$

We say that η is conditionally σ-sub-Gaussian given $\mathcal{F}_\tau\in\mathcal{F}$ for any $\tau\geq 1$ if for all $\lambda\in\mathbb{R}$,

$$\log\mathbb{E}[\exp(\lambda\eta)|\mathcal{F}_\tau] \leq \lambda^2\sigma^2/2.$$

For any $\varepsilon>0$, we denote by $\mathcal{N}(\varepsilon,\mathcal{Z},\|\cdot\|_\infty)$ the ε-covering number of $\mathcal{Z}$ with respect to the supremum norm distance $\|z_1-z_2\|_\infty = \sup_{x\in\mathbb{R}}|z_1(x)-z_2(x)|$. For any $\beta>0$, we define

$$\mathcal{Z}_t(\beta) = \left\{z\in\mathcal{Z} : \sum_{\tau=1}^t (z(X_\tau) - \widehat{z}_t(X_\tau))^2 \leq \beta\right\}.$$

To utilize the concept of Eluder dimension, we introduce the following lemma.

Lemma F.3. *Assume that for any $\tau\geq 1, Y_\tau - z^*(X_\tau)$ is conditionally σ-sub-Gaussian given $\mathcal{F}_{\tau-1}$. Then, for any $\varepsilon>0$ and $\delta\in[0,1]$, with probability at least $1-\delta$, for all $t\geq 1, z^*\in\mathcal{Z}_t(\beta_t(\delta,\varepsilon))$, where*

$$\beta_t(\delta,\varepsilon) = 8\sigma^2\log(\mathcal{N}(\varepsilon,\mathcal{Z},\|\cdot\|_\infty)/\delta) + 4t\varepsilon\left(C + \sqrt{\sigma^2\log(4t(t+1)/\delta)}\right).$$

Proof of Lemma F.3. See Proposition 6 of Russo & Van Roy (2013) for a detailed proof. □

Then we are ready to prove Lemma F.2. It suffices to show $u_h^{(i)} \in \mathcal{U}_h^{k,(i)}$ and $P_h \in \mathcal{P}_h^k$ respectively.

Utility Function Estimation. We denote by $\mathcal{U}$ as the set of all the functions $u : \mathcal{S} \times \mathcal{X}^{(i)} \mapsto [0,1]$ (note that $\mathcal{X}^{(i)} = [0,1]^m$). For any $(k,h,i) \in [K] \times [H] \times [N]$, we set $Y_k^{(i)} = u_h^{k,(i)}$, $X_k^{(i)} = (s_h^k, x_h^{k,(i)})$, and $u^*(\cdot,\cdot) = u_h^{(i)}(\cdot,\cdot)$. We have that $Y_\tau^{(i)} - u^*(X_\tau^{(i)})$ is conditionally $1/2$-sub-Gaussian given $\mathcal{F}_{t(k,h)}$ defined in Definition F.1. By setting

$$\beta^{(1)} = 2\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty) \cdot 4HN/\delta) + 4\left(1 + \sqrt{\log(16K^2HN/\delta)}\right),$$

in Algorithm 1, we can easily check that, using the notion in Lemma F.3, we have that

$$\begin{aligned}\beta^{(1)} &\geq 2\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty) \cdot 4HN/\delta) + 4(k-1)/K \cdot \left(1 + \sqrt{\log(16k(k-1)HN/\delta}\right) \\ &= \beta_{k-1}(\delta/(4HN), 1/K),\end{aligned}$$

for all $k \in [K]$. Thus by Lemma F.3, with probability at least $1 - \delta/(4HN)$, for any $k \in [K]$ we have that

$$u_h^{(i)} = u^\star \in \left\{ u \in \mathcal{U} : \sum_{\tau=1}^k \left(u_h^{k,(i)}(s_h^\tau, x_h^{\tau,(i)}) - u(s_h^\tau, x_h^{\tau,(i)})\right)^2 \leq \beta_{k-1} \right\} \subseteq \mathcal{U}_h^{k,(i)},$$

which gives $u_h^{(i)} \in \mathcal{U}_h^{k,(i)}$ for all $k \in [K]$. Now using a union bound argument over $h \in [H]$ and $i \in [N]$ we conclude that with probability at least $1 - \delta/4$, for all $[K] \times [H] \times [N]$, we have $u_h^{(i)} \in \mathcal{U}_h^{k,(i)}$.

Transition Kernel Estimation. Following Cai et al. (2020b), for any $P \in \mathcal{P}$, we define $z_P : \mathcal{S} \times \mathcal{B} \times [0, HN]^{\mathcal{S}} \to [0, HN]$ by

$$z_P(s, b, f(\cdot)) = \int_{\mathcal{S}} f(s') P(\mathrm{d}s'|s,b), \quad \forall (s, b, f(\cdot)) \in \mathcal{S} \times \mathcal{B} \times [0, HN]^{\mathcal{S}},$$

and $\mathcal{Z} = \{z_P : P \in \mathcal{P}\}$. For any $(k,h) \in [K] \times [H]$, we set $Y_k = \sum_{i=1}^N V_{h+1}^{k,(i)}(s_{h+1}^k)$, $X_k = (s_h^k, b_h^k, \sum_{i=1}^N V_{h+1}^{k,(i)}(\cdot))$, and $z^* = z_{P_h}$. We have that $Y_\tau - z^*(X_\tau)$ is conditionally $HN/2$-sub-Gaussian given $\mathcal{F}_{t(k,h)}$ defined in Definition F.1. Also, by the definition of $\mathcal{P}_h^k$, we have that $\mathcal{Z}_k(\beta) = \left\{z_P : P \in \mathcal{P}_h^k\right\}$. By setting

$$\beta^{(2)} = 2H^2N^2\log(\mathcal{N}(1/(KHN), \mathcal{P}, \|\cdot\|_\infty) \cdot 4H/\delta) + 4HN\left(1 + \sqrt{\log(16K^2H/\delta)}/2)\right),$$

in Algorithm 1, we can check that, using the notion in Lemma F.3, we can show that

$$\begin{aligned}\beta^{(2)} &\geq 2H^2N^2\log(\mathcal{N}(1/K, \mathcal{Z}, \|\cdot\|_\infty) \cdot 4H/\delta) + 4(k-1)/K \cdot \left(HN + \sqrt{H^2N^2/4 \cdot \log(16k(k-1)H/\delta)}\right) \\ &= \beta_{k-1}(\delta/4H, 1/K)\end{aligned}$$

for all $k \in [K]$, where we leave the proof of $\mathcal{N}(\varepsilon, \mathcal{Z}, \|\cdot\|_\infty) \leq \mathcal{N}(\varepsilon/HN, \mathcal{P}, \|\cdot\|_{\infty,1})$ in the end. Applying Lemma F.3, with probability at least $1 - \delta/4H$, for all $k \in [K]$, we have that

$$z^* \in \mathcal{Z}_k(\beta_{k-1}(\delta/4H, 1/K)) \subset \mathcal{Z}_k(\beta),$$

which implies that $P_h \in \mathcal{P}_h^k$. Now applying a union bound over all $h \in [H]$, with probability at least $1 - \delta/4$, for all $(k,H) \in [K] \times [H]$, we have that $P_h \in \mathcal{P}_h^k$. In the sequel, we show that $\mathcal{N}(\varepsilon, \mathcal{Z}, \|\cdot\|_\infty) \leq \mathcal{N}(\varepsilon/HN, \mathcal{P}, \|\cdot\|_{\infty,1})$ for any $\varepsilon > 0$. Indeed, this is proved by using the fact that for any $z_P, z_{P'} \in \mathcal{Z}$ with $P, P' \in \mathcal{P}$, we have that

$$\begin{aligned}\|z_P - z_{P'}\|_\infty &= \sup_{(s,b,f(\cdot)) \in \mathcal{S}\times\mathcal{B}\times[0,HN]^S} \left| \int_{\mathcal{S}} f(s') P(\mathrm{d}s'|s,b)\mathrm{d}s' - \int_{\mathcal{S}} V(s') P'(s'|s,b)\mathrm{d}s' \right| \\ &\leq \sup_{(s,b,f(\cdot)) \in \mathcal{S}\times\mathcal{B}\times[0,HN]^{\mathcal{S}}} HN \cdot \int_{\mathcal{S}} |P(\mathrm{d}s'|s,b) - P'(\mathrm{d}s'|s,b)| = HN \cdot \|P - P'\|_{\infty,1}.\end{aligned}$$

Thus we have proved the results for both utility function and transition kernel. Now applying a union bound we can conclude that with the choice of $\beta^{(1)}$, $\beta^{(2)}$ in Lemma F.2, with probability at least $1 - \delta/2$, event E holds. □

Now based on event E, we present the proof of the two conclusions in Lemma C.3 respectively.

Conclusion 1: Optimism. The optimism result directly holds under event E. In fact, for any $(k,h) \in [K] \times [H]$, when $P_h \in \mathcal{P}_h^k$ and $u_h^{(i)} \in \mathcal{U}_h^{k,(i)}$, by the definition of $Q_h^{k,(i)}$ it holds that

$$\begin{aligned}
-\iota_h^k(s_h,a_h,b_h) &= \sum_{i=1}^N \max_{u\in\mathcal{U}_h^{k,(i)}} u(s_h,x_h^{(i)}) - u_h(s_h,x_h^{(i)}) \\
&\qquad + \sum_{i=1}^N Q_h^{k,(i)}(s_h,x_h^{(i)}) - \widehat{u}_h^{(i)}(c_h,x_h^{(i)}) - \int_{\mathcal{S}} P_h(\mathrm{d}s'|s_h,b_h)V_{h+1}^{k,(i)}(s') \\
&\geq \sum_{i=1}^N Q_h^{k,(i)}(s_h,x_h^{(i)}) - \widehat{u}_h^{(i)}(c_h,x_h^{(i)}) - \int_{\mathcal{S}} P_h(\mathrm{d}s'|s_h,b_h)V_{h+1}^{k,(i)}(s') \\
&= \sum_{i=1}^N \mathtt{Clip}_{[0,H-h]}\left(\int_S \widehat{P}_h^k(\mathrm{d}s'|s_h,b_h)V_{h+1}^{k,(i)}(s')\right) - \int_{\mathcal{S}} P_{h+1}(\mathrm{d}s'|s_h,b_h)V_{h+1}^{k,(i)}(s'),
\end{aligned} \tag{54}$$

where the inequality follows from event E and the optimistic choice of $\widehat{u}_h^{k,(i)}$. Also, for $h=H$, the right-hand side of (54) is zero since $V_{H+1}^{k,(i)}(\cdot)=0$. For $h<H$, by the construction of $Q_{h+1}^{k,(i)}$ and the fact that $\widehat{u}_{h+1}^{k,(i)}(\cdot,\cdot) \in [0,1]$,

$$Q_{h+1}^{k,(i)}(s_{h+1},x_{h+1}^{(i)},b_{h+1}) \in [0,H-h], \quad V_{h+1}^{k,(i)}(s_{h+1}) \in [0,H-h], \quad \int_{\mathcal{S}} V_{h+1}^{k,(i)}(s')\, P_h(\mathrm{d}s'|s_h,b_h) \in [0,H-h].$$

For any $s_h, s_{h+1} \in \mathcal{S}$, $a_{h+1} \in \mathcal{A}$, and $b_h, b_{h+1} \in \mathcal{B}$. Thus, it yields from (54) that

$$\begin{aligned}
-\iota_h^k(s_h,a_h,b_h) &= \sum_{i=1}^N \int_S \widehat{P}_h^k(\mathrm{d}s'|s_h,b_h)V_{h+1}^{k,(i)}(s') - \int_{\mathcal{S}} P_{h+1}(\mathrm{d}s'|s_h,b_h)V_{h+1}^{k,(i)}(s') \\
&= \max_{P\in\mathcal{P}_h^k} \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{k,(i)}(s')P(\mathrm{d}s'|s_h,b_h) - \int_{\mathcal{S}} \sum_{i=1}^N V_{h+1}^{k,(i)}(s')P_h(\mathrm{d}s'|s_h,b_h) \geq 0,
\end{aligned} \tag{55}$$

where the last step follows from the definition of event E and the optimistic choice of $\widehat{P}_h^k$. Therefore,

$$\iota_h^k(s_h,a_h,b_h) = \sum_{i=1}^N u_h^{k,(i)}(s_h,x_h^{(i)}) + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s_h,b_h)V_{h+1}^{k,(i)}(s') - Q_h^{k,(i)}(s_h,x_h^{(i)},b_h) \leq 0,$$

for all $(s_h, x_h^{(i)}, b_h) \in \mathcal{S} \times \mathcal{X}^{(i)} \times \mathcal{B}$. This proves the first conclusion of Lemma C.3.

Conclusion 2: Accuracy. First we introduce the following lemma to decompose the martingale.

Lemma F.4 (Martingale Decomposition)**.** *For any $k \in [K]$, we have the following decomposition,*

$$\sum_{i=1}^N V_1^{k,(i)}\left(s_1^k\right) - V_1^{(\pi^k,\nu^k),(i)}\left(s_1^k\right) = \sum_{h=1}^H D_h^k + \sum_{h=1}^H -\iota_h^k(s_h^k,a_h^k,b_h^k),$$

where ι_h^k is defined in (38) and the term D_h^k takes the form

$$D_h^k = \int_{\mathcal{S}} \sum_{i=1}^N \left(V_{h+1}^{k,(i)}(s') - V_{h+1}^{(\pi^k,\nu^k),(i)}(s')\right) P_h(\mathrm{d}s'|s_h^k,a_h^k) - \sum_{i=1}^N V_{h+1}^{k,(i)}(s_{h+1}^k) - V_{h+1}^{(\pi^k,\nu^k),(i)}(s_{h+1}^k)$$

Moreover, we have $D_H^k = 0$ for all $k \in [K]$, and $D_1^1, D_2^1, D_3^1, \cdots, D_{H-1}^1, D_1^2, D_2^2, \cdots$, is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 1}$ defined in Definition F.1, where each term is bounded by $2HN$.

Proof of Lemma F.4. See Lemma F.1 of Cai et al. (2020b) for a detailed proof. □

Upon applying Lemma F.4, we have the following decomposition:

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{N} V_1^{k,(i)}(s_1^k) - V_1^{(\pi^k,\nu^k),(i)}(s_1^k) = &\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} D_h^k}_{\text{(i)}} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{N}\left(-u_h^{(i)}(s_h^k, x_h^{k,(i)}) + \widehat{u}_h^{k,(i)}(s_h^k, x_h^{k,(i)})\right)}_{\text{(ii)}} \\ &+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{i=1}^{N}\left(Q_h^{k,(i)}(s_h^k, x_h^{k,(i)}, b_h^k) - \widehat{u}_h^{k,(i)}(s_h^k, x_h^{k,(i)}) - \int_{\mathcal{S}} P_h(\mathrm{d}s'|s_h^k, b_h^k) V_{h+1}^{k,(i)}(s')\right)}_{\text{(iii)}}.\end{aligned} \tag{56}$$

For term (i) in (56), note that $|D_h^k| \le 2NH$, $D_H^k = 0$ for all $(k,h) \in [K]\times[H]$, and

$$D_1^1, D_2^1, D_3^1, \cdots, D_{H-1}^1, D_1^2, D_2^2, \cdots$$

is a martingale difference sequence. Using the Azuma-Hoeffding inequality, we obtain that, with probability at least $1-\delta/2$, it holds that

$$\text{(i)} \le \sqrt{8KH^3N^2\log(2/\delta)}. \tag{57}$$

To deal with the remaining two terms, we introduce the following lemma.

Lemma F.5 (Telescoping Sum)**.** *For any $\alpha > 0$, and $\beta > 0$, we have*

$$\sum_{k=1}^{K} \sup_{z,z'\in\mathcal{Z}_k(\beta)} |z(x_k) - z'(x_k)| \le 1 + C\cdot d + 4\cdot\sqrt{d\beta K}.$$

where $d = \dim_{\mathrm{E}}(\mathcal{Z}, 1/K)$.

Proof of Lemma F.5. See Lemma 5 of Russo & Van Roy (2013) for a detailed proof. □

For term (ii) in (56), under event E and applying Lemma F.5, it holds that, with $d_1 := \dim_{\mathbb{E}}(\mathcal{U}, 1/K)$,

$$\text{(ii)} \le HN\sum_{k=1}^{K} \max_{u\in\mathcal{U}_h^{k,(i)}} u(s_h^k, x_h^{k,(i)}) - \min_{u\in\mathcal{U}_h^{k,(i)}} u(s_h^k, x_h^{k,(i)}) \le HN(1 + d_1 + 4\sqrt{d_1\beta^{(1)}K}), \tag{58}$$

For term (iii) in (56), under event E and applying Lemma F.5, it holds that, with $d_2 := \dim_{\mathbb{E}}(\mathcal{Z}_P, 1/K)$,

$$\begin{aligned}\text{(iii)} &\le H\sum_{k=1}^{K}\max_{P\in\mathcal{P}_h^k}\int_{\mathcal{S}}\sum_{i=1}^{N} V_{h+1}^{k,(i)}(s')P(\mathrm{d}s'|s_h^k, b_h^k) - \int_{\mathcal{S}}\sum_{i=1}^{N} V_{h+1}^{k,(i)}(s')P_h(\mathrm{d}s'|s_h^k, b_h^k) \\ &\le H\sum_{k=1}^{K}\max_{P\in\mathcal{P}_h^k}\int_{\mathcal{S}}\sum_{i=1}^{N} V_{h+1}^{k,(i)}(s')P(\mathrm{d}s'|s_h^k, b_h^k) - \min_{P\in\mathcal{P}_h^k}\int_{\mathcal{S}}\sum_{i=1}^{N} V_{h+1}^{k,(i)}(s')P(\mathrm{d}s'|s_h^k, b_h^k) \\ &\le H(1 + d_1HN + 4\sqrt{d_2\beta^{(2)}K}),\end{aligned} \tag{59}$$

Finally, combining bounds (57), (58), and (59), we conclude that with probability at least $1-\delta$,

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{i=1}^{N} V_1^{k,(i)}(s_1^k) - V_1^{(\pi^k,\nu^k),(i)}(s_1^k) \\ &\quad\le \sqrt{8KH^3N^2\log(2/\delta)} + HN(1 + d_1 + 4\sqrt{d_1\beta^{(1)}K}) + H(1 + d_1HN + 4\sqrt{d_2\beta^{(2)}K}) \\ &\quad\le \sqrt{8KH^3N^2\log(2/\delta)} + (1+d)H(H+N) + 4HN\sqrt{d\beta^{(1)}K} + 4\sqrt{d\beta^{(2)}K} \\ &\quad\le \sqrt{8KH^3N^2\log(2/\delta)} + (1+d)H(H+N) + 4H\sqrt{2d(N^2\beta^{(1)} + \beta^{(2)})K},\end{aligned} \tag{60}$$

where $d = \max\{d_1, d_2\}$ and the last inequality follows from the fact that the inequality $\sqrt{x} + \sqrt{y} \le \sqrt{2(x+y)}$. This proves the second conclusion in Lemma C.3 and finishes the proof of Lemma C.3. □

F.4. Proof for Theorem 3.4: Regret for Fair Division Property

Proof. Recalling the definition of FD loss in (14), note that

$$\begin{aligned}\ell_h^{\mathrm{PE}}(\nu^k, s_h) &= \inf_{x\in \mathrm{SI}(s_h,h)} \sum_{i=1}^N (u_h^{(i)}(s_h, x^{(i)}) - u_h^{(i)}(s_h, \nu_h^{k,(i)}(s_h)))\\ &\le \sum_{i=1}^N (u_h^{(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h)) - u_h^{(i)}(s_h, \nu_h^{k,(i)}(s_h))).\end{aligned}$$

Also, observe that

$$\begin{aligned}\ell_h^{\mathrm{SI}}(\nu^k, s_h) &= \sum_{i=1}^N (u_h^{(i)}(s_h, e^{(i)}) - u_h^{(i)}(s_h, \nu_h^{k,(i)}(s_h)))\\ &\le \sum_{i=1}^N (u_h^{(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h)) - u_h^{(i)}(s_h, \nu_h^{k,(i)}(s_h))),\end{aligned}$$

where the inequality originates from the definition of $\nu_h^\star$ in Definition 2.2. Under the event E defined in (53), following the same procedure as in dealing with term (ii) in (56), it holds for all $h \in [H]$ that,

$$\begin{aligned}\sum_{k=1}^K \mathbb{E}_{\pi^k} \ell_h^{\mathrm{FD}}(\nu^k, s_h) &\le \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}_{\pi^k}\Big[e\Big(u_h^{(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h)) - \widehat{u}_h^{(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h))\Big)\\ &\qquad + \Big(\widehat{u}_h^{(i)}(s_h, \nu_h^{\star,(i)}(\nu^k)(s_h)) - \widehat{u}_h^{(i)}(s_h, \nu_h^{k,(i)}(s_h))\Big)\Big]\\ &\qquad + \Big(\widehat{u}_h^{(i)}(s_h, \nu_h^{k,(i)}(s_h)) - u_h^{(i)}(s_h, \nu_h^{k,(i)}(s_h))\Big)\\ &\le \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}_{\pi^k}\left[\max_{u\in \mathcal{U}_h^{k,(i)}} u(s_h, x_h^{(i)}) - \min_{u\in \mathcal{U}_h^{k,(i)}} u(s_h, x_h^{(i)})\right]\\ &\le N((1+d_1)N + N\sqrt{d\beta^{(1)}K}),\end{aligned}$$

where we remark that the first two terms in the first line are non-positive because of the definition of event E and ν^k. Applying the definition of regret for fair division property $\mathrm{Regret}_{\mathrm{FD}}(K)$, we have that

$$\mathrm{Regret}_{\mathrm{FD}}(K) \le \sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\pi^k} \ell_h^{FD}(\nu^k, s_h^k) \le H((1+d_1)N + N\sqrt{d\beta^{(1)}K}) \le \mathcal{O}(\sqrt{dH^2N^2\beta^{(1)}K}).$$

This finishes the proof of Theorem 3.4 on the regret for fair division. □

G. Proofs of Offline Learning Algorithm: Section 4

G.1. Proof of Lemma C.4

Proof of Lemma C.4. By the definition of suboptimality in (11) and Lemma C.2, it holds that

$$\begin{aligned}\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}) &= \sum_{i=1}^N \big[V_{1,\widehat{u},\widehat{P}}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu})),(i)} - \widehat{V}_{1,\widehat{u},\widehat{P}}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu})),(i)} + \widehat{V}_{1,\widehat{u},\widehat{P}}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu})),(i)} - V_1^{(\widehat{\pi},\widehat{\nu}),(i)}\big](s_1)\\ &\le \sum_{i=1}^N \big[V_1^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu})),(i)} - \widehat{V}_{1,\widehat{u},\widehat{P}}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu})),(i)} + \widehat{V}_{1,\widehat{u},\widehat{P}}^{(\widehat{\pi},\widehat{\nu}),(i)} - V_1^{(\widehat{\pi},\widehat{\nu}),(i)}\big](s_1).\end{aligned} \tag{61}$$

For notational simplicity, we define $\Delta_h(s) := \sum_{i=1}^N \big[V_1^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu})),(i)} - \widehat{V}_{1,\widehat{u},\widehat{P}}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu})),(i)}\big](s)$ and abbreviate $d_h^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}$ as d_h (Recall the definition of $d_h^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}$ in (28)). To proceed further, we define that

$$\mu_h^{\widehat{u}} := \mathbb{E}_{d_h} \sum_{i=1}^N |\widehat{u}_h^{(i)}(s, x^{(i)}) - u_h(s, x^{(i)})| \text{ and } \mu_h^{\widehat{P}} := \mathbb{E}_{d_h} \|\widehat{P}_h(\cdot \mid s, b) - P_h(\cdot \mid s, b)\|_1. \tag{62}$$

By the definition of $\widehat{V}^{(\pi,\nu),(i)}_{h,\widehat{u},\widehat{P}}$, $\mu^{\widehat{u}}_h$ and $\mu^{\widehat{P}}_h$, it yields that

$$\begin{aligned}\mathbb{E}_{d_h}\Delta_h(s_h) &= \mathbb{E}_{d_h}\sum_{i=1}^{N}\big[(P_hV^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}_{h+1} - P_h\widehat{V}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}_{h+1} + P_h\widehat{V}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}_{h+1} - \widehat{P}_h\widehat{V}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}_{h+1,(\widehat{u},\widehat{P})})(s_h,b_h)\\ &\quad + (u^{(i)}_h - \widehat{u}^{(i)}_h)(s_h,x^{(i)}_h)\big]\\ &\le \mathbb{E}_{d_h}\sum_{i=1}^{N}\big[\int_{\mathcal{S}}(V^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}_{h+1} - \widehat{V}^{(\pi^\dagger(\widehat{\nu}),\nu^\star(\widehat{\nu}))}_{h+1,(\widehat{u},\widehat{P})})P_h(\mathrm{d}s_{h+1}\,|\,s_h,b_h)\big] + HN\mu^{\widehat{P}}_h + \mu^{\widehat{u}}_h\\ &= \mathbb{E}_{d_{h+1}}\Delta_{h+1}(s_{h+1}) + HN\mu^{\widehat{P}}_h + \mu^{\widehat{u}}_h\end{aligned} \tag{63}$$

where the inequality is based on the fact that both the true and estimated value functions are bounded by $[0,H]$. Hence, by telescoping index h over $[H]$, it holds that

$$\begin{aligned}\Delta_1(s_1) &= \sum_{h=1}^{H} HN\mu^{\widehat{P}}_h + \mu^{\widehat{u}}_h \le \sum_{h=1}^{H}\sqrt{\mathbb{E}_{\rho_h}\left(\frac{d_h}{\rho_h}\right)^2}\big(HN\sqrt{\epsilon^{\widehat{P}}_h} + \sqrt{\epsilon^{\widehat{u}}_h}\big)\\ &\le \sqrt{C^\star_\rho}\big(\sum_{h=1}^{H} HN\sqrt{\epsilon^{\widehat{P}}_h} + \sqrt{\epsilon^{\widehat{u}}_h}\big),\end{aligned}$$

where the first inequality relies on Cauchy-Schwarz inequality and the definition of $\epsilon^{\widehat{P}}_h$ and $\epsilon^{\widehat{u}}_h$ in Lemma C.4, and the second inequality relies on the definition of $C^\star_\rho$ in (30). Plugging $\Delta_1(s_1)$ into (61), we conclude the proof of Lemma C.4. □

G.2. Proof of Lemma C.5

Proof of Lemma C.5. According to the choice of $\widehat{P}$ in Algorithm 1, we first have that

$$\sum_{i=1}^{N}\widehat{V}^{(\widehat{\pi},\widehat{\nu}),(i)}_{1,(\widehat{P},\widehat{u})}(s_1) - V^{(\widehat{\pi},\widehat{\nu}),(i)}_1(s_1) \le \sum_{i=1}^{N}\widehat{V}^{(\widehat{\pi},\widehat{\nu}),(i)}_{1,(P,\widehat{u})}(s_1) - \sum_{i=1}^{N}V^{(\widehat{\pi},\widehat{\nu}),(i)}_1(s_1), \tag{64}$$

since $\widehat{P} = \{\widehat{P}_h\}_{h\in[H]}$ is the global pessimistic estimator in $\mathcal{P}_{h,\xi_2}$ and $P_h \in \mathcal{P}_{h,\xi_2}$. Next, we show by induction that the right-hand side of (64) is non-positive. For step $h = H$, we have that

$$\sum_{i=1}^{N}\widehat{V}^{(\widehat{\pi},\widehat{\nu}),(i)}_{H,(P,\widehat{u})}(s_H) - \sum_{i=1}^{N}V^{(\widehat{\pi},\widehat{\nu}),(i)}_H(s_H) = \sum_{i=1}^{N}\widehat{u}^{(i)}_H(s_H,\widehat{\nu}^{(i)}_H(s_H)) - \sum_{i=1}^{N}u^{(i)}_H(s_H,\widehat{\nu}^{(i)}_H(s_H)) \le 0, \quad \forall s_H\in\mathcal{S},$$

since $u^{(i)}_H \in \mathcal{U}^{(i)}_{H,\xi_1}$ and $\widehat{u}^{(i)}_H$ is pessimistic estimator for each $i\in[N]$. Now suppose that inequality

$$\sum_{i=1}^{N}\widehat{V}^{(\widehat{\pi},\widehat{\nu}),(i)}_{h+1,(P,\widehat{u})}(s_{h+1}) - \sum_{i=1}^{N}V^{(\widehat{\pi},\widehat{\nu}),(i)}_{h+1}(s_{h+1}) \le 0, \quad \forall s_{h+1}\in\mathcal{S},$$

holds for step $h+1$. Then for step h, we have that

$$\begin{aligned}\sum_{i=1}^{N}\widehat{V}^{(\widehat{\pi},\widehat{\nu}),(i)}_{h,(P,\widehat{u})}(s_h) - \sum_{i=1}^{N}V^{(\widehat{\pi},\widehat{\nu}),(i)}_h(s_h) &= \underbrace{\sum_{i=1}^{N}\widehat{u}^{(i)}_h(s_h,\widehat{\nu}^{(i)}_h(s_h)) - \sum_{i=1}^{N}u^{(i)}_h(s_h,\widehat{\nu}^{(i)}_h(s_h))}_{\text{(i)}}\\ &\quad + \underbrace{\int_{\mathcal{S}}\left(\sum_{i=1}^{N}\widehat{V}^{(\widehat{\pi},\widehat{\nu}),(i)}_{h+1,(P,\widehat{u})}(s') - \sum_{i=1}^{N}V^{(\widehat{\pi},\widehat{\nu}),(i)}_{h+1}(s')\right)P_h(\mathrm{d}s'|s_h,\widehat{\pi}_h(s_h))}_{\text{(ii)}}, \quad \forall s_h\in\mathcal{S},\end{aligned}$$

where (i) ≤ 0 relies on the fact that $u^{(i)}_h \in \mathcal{U}^{(i)}_{h,\xi_1}$ and $\widehat{u}^{(i)}_h$ is pessimistic estimator for each $i\in[N]$. By induction, we prove that (ii) ≤ 0. Thus we conclude that the right-hand side of (64) is non-positive. □

G.3. Missing Proofs of Theorem C.6

Before we start the proof of Theorem C.6, we introduce the following lemma, which plays the key role in sharpening the convergence rate in the analysis for both estimated kernels and utility functions.

Lemma G.1 (Uniform Bernstein Inequality with Covering Number)**.** *For any given functional class $\mathcal{F} \subset \{f : \mathcal{X} \mapsto \mathbb{R}\}$, where $\mathcal{X}$ is a probability space. If we assume that the ϵ-covering number of $\mathcal{F}$ under infinity-norm is finite, that is, $M := \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty) < \infty$ and we also assume that there exists an absolute constant such that $|f(X)| \le R$ a.s., then the following inequality holds for all $f \in \mathcal{F}$ with probability at least $1-\delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right| \le 2\epsilon + \sqrt{\frac{2\mathbb{V}[f(X)]\log(M/\delta)}{n}} + 4\sqrt{\frac{R\epsilon\log(M/\delta)}{n}} + \frac{2R\log(M/\delta)}{3n},$$

where $X, X_1, ..., X_n$ are all i.i.d. samples on the probability space $\mathcal{X}$.

Proof of Lemma G.1. To obtain this lemma, we adapted Bernstein inequality with the technique dealing with covering number. See Appendix H.1 for detailed proof. □

Proof of Theorem C.6. We prove the theorem by the following lemmas, which are adapted from Xie et al. (2021) are based on Lemma G.1. For notational simplicity, we define $\|f\|_{2,\rho}$ as $\sqrt{\mathbb{E}_\rho[|f|^2]}$.

Lemma G.2. *For any $(i,h) \in [N] \times [H]$, it holds with probability at least $1-\delta/NH$ that any $\widehat{u} \in \mathcal{U}_{h,\xi_1}^{(i)}$ satisfies*

$$\left|\left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h} - \left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}\right| \le \sqrt{\frac{82\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}}.$$

Proof of Lemma G.2. See Appendix H.2 for a detailed proof. □

Lemma G.3. *For any $(i,h) \in [N] \times [H]$, it holds with probability at least $1-\delta/NH$ that any $\widehat{u}, \widetilde{u} \in \mathcal{U}_{h,\xi_1}^{(i)}$ satisfy*

$$\begin{aligned}
&\Big|\left\|\widehat{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}^2 - \left\|\widetilde{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}^2 \\
&\quad - \left\|\widehat{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}^2 + \left\|\widetilde{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}^2\Big| \\
&\le \left\|\widehat{u}(s,x^{(i)}) - \widetilde{u}(s,x^{(i)})\right\|_{2,\rho_h} \cdot \sqrt{\frac{64\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}} + \frac{262\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K}.
\end{aligned}$$

Proof of Lemma G.3. See Appendix H.3 for a detailed proof. □

Lemma G.4 (Concentration)**.** *By setting*

$$\xi_1 = \frac{\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)\cdot NH/\delta)}{K},$$

it holds with probability at least $1-\delta/4$ that for any $(i,h) \in [N] \times [H]$, $u_h^{(i)} \in \mathcal{U}_{h,\xi_1}^{(i)}$.

Proof of Lemma G.4. This is a trivial conclusion since we note that $u_h^{(i)}(s_h^\tau, x_h^{\tau,(i)}) = u_h^{\tau,(i)}$. □

Lemma G.5 (Accuracy)**.** *It holds with probability at least $1-\delta/4$ that, for any $(i,h) \in [N] \times [H]$ and $u \in \mathcal{U}_{h,\xi_1}^{(i)}$,*

$$\left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}^2 \le \frac{225\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)\cdot NH/\delta)}{K}.$$

Proof of Lemma G.5. By Lemma G.3 with $\widehat{u} = u$ and $\widetilde{u} = u_h^{(i)}$, it holds that with probability at least $1 - \delta/4NH$, for any $u \in \mathcal{U}$,

$$\begin{aligned}
&\left\|u(s, x^{(i)}) - u_h^{(i)}(s, x^{(i)})\right\|_{2,\rho_h}^2 \\
&\quad \le \frac{1}{K}\sum_{\tau=1}^{K}\left(u(s_h^\tau, x_h^{\tau,(i)}) - u_h^{(i)}(s_h^\tau, x_h^{\tau,(i)})\right)^2 + \frac{262\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)4NH/\delta)}{3K} \\
&\qquad + \left\|\widehat{u}(s, x^{(i)}) - u_h^{(i)}(s, x^{(i)})\right\|_{2,\rho_h} \cdot \sqrt{\frac{64\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)4NH/\delta)}{K}}.
\end{aligned}$$

Now for any $(i, h) \in [N] \times [H]$, we restrict $u \in \mathcal{U}_{h,\xi_1}^{(i)}$ to obtain that

$$\begin{aligned}
&\left\|u(s, x^{(i)}) - u_h^{(i)}(s, x^{(i)})\right\|_{2,\rho_h}^2 \\
&\quad \le \xi_1 + \left\|\widehat{u}(s, x^{(i)}) - u_h^{(i)}(s, x^{(i)})\right\|_{2,\rho_h} \cdot \sqrt{\frac{64\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)4NH/\delta)}{K}} \\
&\qquad + \frac{262\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)4NH/\delta)}{3K} \\
&\quad \le \left\|\widehat{u}(s, x^{(i)}) - u_h^{(i)}(s, x^{(i)})\right\|_{2,\rho_h} \cdot \sqrt{\frac{64\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)4NH/\delta)}{K}} \\
&\qquad + \frac{90\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)4NH/\delta)}{K}.
\end{aligned} \tag{65}$$

Solving the quadratic inequality in (65), we have that

$$\left\|u(s, x^{(i)}) - u_h^{(i)}(s, x^{(i)})\right\|_{2,\rho_h} \le \sqrt{\frac{225\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty)4NH/\delta)}{K}}.$$

Finally, applying a union bound argument over $(i, h) \in [N] \times [H]$, we finish the proof of Lemma G.5. □

This proves that with probability at least $1 - \delta/4$, we have $\epsilon_h^{\widehat{u}} \le 225\log(\mathcal{N}(1/K, \mathcal{U}, \|\cdot\|_\infty) \cdot 4NH/\delta)/K$. Combining this result with Lemma G.4, we finishes the proof of Theorem C.6. □

G.4. Missing Proofs of Theorem C.7

Proof of Theorem C.7. As the first part of Theorem C.7, we introduce the the following key lemma to show that the event

$$E_1 := \left\{\mathbb{E}_{\mathcal{D}_h}\|\widehat{P}_h^{\mathrm{MLE}}(\cdot \mid s, b) - P_h(\cdot \mid s, b)\|_1^2 \le C'\log(\mathcal{N}_{[]}(1/K, \mathcal{P}, \|\cdot\|_{1,\infty})H/\delta)/K, \text{ for all } h \in [H]\right\}$$

happens with probability at least $1 - \delta/4$, where C' is an absolute constant.

Lemma G.6. *According to Algorithm 2 , then event E_1 happens with probability at least $1 - \delta/4$.*

Proof of Lemma G.6. For the simiplicity of notation, we denote by $g_h(\widehat{P})(s, b) := \|\widehat{P}(\cdot \mid s, b) - P_h(\cdot \mid s, b)\|_1^2$. By the following lemma, we show that MLE estimation in Algorithm 2) can converge at a negative square root rate.

Lemma G.7 (MLE Estimation Guarantee)**.** *According to Algorithm 2 , then event*

$$E_2 := \left\{\mathbb{E}_{\rho_h} g_h(\widehat{P}_h^{\mathrm{MLE}}) \le c'\log(\mathcal{N}_{[]}(1/K, \mathcal{P}, \|\cdot\|_{1,\infty})H/\delta)/K, \textit{ for all } h \in [H]\right\}$$

happens with probability at least $1 - \delta/8$, where c' is an absolute constant.

Proof of Lemma G.7. See Appendix H.3 for detailed proof. □

Notice that the gap between Lemma G.6 and Lemma G.7 can be bridged by concentration analysis which relies on the adapted Bernstein inequality in Lemma G.1. We introduce the following lemma.

Lemma G.8 (Bernstein Inequality with Union Bound I)**.** *According to Algorithm 2, if we define the event*

$$E_3 := \Big\{ \big| [\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h}] g_h(\widehat{P}_h^{\mathrm{MLE}}) \big| \le c'' \log(\mathcal{N}_{[]}(1/K, \mathcal{P}, \|\cdot\|_{1,\infty}) H/\delta)/K, \text{ for all } h \in [H] \Big\},$$

then $E_2 \cap E_3$ happens with probability at least $1 - \delta/8$, where c'' is an absolute constant.

Proof of Lemma G.8. See Appendix H.3 for detailed proof. □

Since it holds that

$$\mathbb{E}_{\rho_h} g_h(\widehat{P}_h^{\mathrm{MLE}}) \le \mathbb{E}_{\mathcal{D}_h} g_h(\widehat{P}_h^{\mathrm{MLE}}) + [\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h}] g_h(\widehat{P}_h^{\mathrm{MLE}}),$$

Lemma G.6 is the direct consequence of Lemma G.7 and Lemma G.8. □

With Lemma G.6, the last part of the proof of Theorem C.7 is to upper bound $\sup_{\in[H]} \epsilon_h^{\widehat{P}}$. Recall that we denote by $g_h(\widehat{P})(s,b) := \|\widehat{P}(\cdot \,|\, s,b) - P_h(\cdot \,|\, s,b)\|_1^2$. On the event E_1, we decompose $\epsilon_h^{\widehat{P}}$ as follows.

$$\begin{aligned}
\epsilon_h^{\widehat{P}} &= \mathbb{E}_{\mathcal{D}_h} g_h(\widehat{P}_h^{\mathrm{MLE}}) + [\mathbb{E}_{\rho_h} - \mathbb{E}_{\mathcal{D}_h}] g_h(\widehat{P}_h) + \mathbb{E}_{\mathcal{D}_h}(g_h(\widehat{P}) - g_h(\widehat{P}_h^{\mathrm{MLE}})) \\
&\le 2\mathbb{E}_{\mathcal{D}_h} g_h(\widehat{P}_h^{\mathrm{MLE}}) + [\mathbb{E}_{\rho_h} - \mathbb{E}_{\mathcal{D}_h}] g_h(\widehat{P}_h) + 2\mathbb{E}_{\mathcal{D}_h} \|\widehat{P}_h(\cdot \,|\, s,b) - \widehat{P}_h^{\mathrm{MLE}}(\cdot \,|\, , s, b)\|_1^2 \\
&\le 2\mathbb{E}_{\mathcal{D}_h} g_h(\widehat{P}_h^{\mathrm{MLE}}) + [\mathbb{E}_{\rho_h} - \mathbb{E}_{\mathcal{D}_h}] g_h(\widehat{P}_h) + 2\xi_2 \\
&\le 4\xi_2 + [\mathbb{E}_{\rho_h} - \mathbb{E}_{\mathcal{D}_h}] g_h(\widehat{P}_h),
\end{aligned} \tag{66}$$

where the first inequality relies on the fact that $(a+b)^2 \le 2(a^2+b^2)$ and the last inequality relies on the definition of E_1. Hence it suffices to upper bound the second term in (66). Motivated by Uehara & Sun (2021) and the proof of Lemma G.2, we prove the following lemma based on the adapted Bernstein inequality in Lemma G.1.

Lemma G.9 (Bernstein Inequality with Union Bound II)**.** *According to Algorithm 2 and selecting $\xi_2 = C' \log(\mathcal{N}_{[]}(1/K, \mathcal{P}, \|\cdot\|_{1,\infty})/\delta)/K$, if we define the event*

$$E_4 := \big\{ \big| [\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h}] g_h(\widehat{P}_h) \big| \le C'' \log(\mathcal{N}_{[]}(1/K, \mathcal{P}, \|\cdot\|_{1,\infty}) H/\delta)/K, \text{ for all } h \in [H] \big\},$$

then $E_1 \cap E_4$ happens with probability at least $1 - \delta/8$, where C'' is an absolute constant.

Proof. See Appendix H.3 for detailed proof. □

Apply Lemma H.3 and Lemma G.6. Based on $E_1 \cap E_2 \cap E_3$ and the selection of ξ_2, then

$$\sup_{h \in [H]} \epsilon_h^{\widehat{P}} \le (c' + c'' + C'') \log(\mathcal{N}_{[]}(1/K, \mathcal{P}, \|\cdot\|_{1,\infty}) H/\delta)/K,$$

which concludes the proof for Theorem C.7. □

G.5. Proof for Theorem 4.4: Offline Fair Division Loss

Proof. Similar to the proof for Theorem 3.4, the following two inequalities originate from the definition of $\nu_h^\star$ in (4).

$$\begin{aligned}
\ell_h^{\mathrm{PE}}(\widehat{\nu}, s_h) &= \inf_{x \in \mathcal{PE}(s_h, h)} \sum_{i=1}^N \Big(u_h^{(i)}(s_h, x^{(i)}) - u_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h)) \Big) \\
&\le \sum_{i=1}^N \Big(u_h^{(i)}(s_h, \nu_h^{\star,(i)}(\widehat{\nu})(s_h)) - u_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h)) \Big).
\end{aligned}$$

Also, observe that

$$\begin{aligned}
\ell_h^{\mathrm{SI}}(\widehat{\nu}, s_h) &= \sum_{i=1}^N \Big(u_h^{(i)}(s_h, e^{(i)}) - u_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h)) \Big) \\
&\le \sum_{i=1}^N \Big(u_h^{(i)}(s_h, \nu_h^{\star,(i)}(\widehat{\nu})(s_h)) - u_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h)) \Big),
\end{aligned}$$

By the definition of offline FD loss defined in (24), it holds that

$$\begin{aligned}\mathcal{L}_{\mathrm{FD}} \leq \sum_{i=1}^{N}\sum_{h=1}^{H} \mathbb{E}_{\rho_h}\Big(u_h^{(i)}(s_h, \nu_h^{\star,(i)}(\widehat{\nu})(s_h)) - \widehat{u}_h^{(i)}(s_h, \nu_h^{\star,(i)}(\widehat{\nu})(s_h))\Big) \\ + \Big(\widehat{u}_h^{(i)}(s_h, \nu_h^{\star,(i)}(\widehat{\nu})(s_h)) - \widehat{u}_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h))\Big) \\ + \Big(\widehat{u}_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h)) - u_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h))\Big)\end{aligned}$$

By the definition of $\widehat{\nu}$ and the event E_0 defined in and (C.6) defined in Theorem C.6, it further holds with at probability at least $1-\delta$ that

$$\begin{aligned}\mathcal{L}_{\mathrm{FD}} &\leq \sum_{i=1}^{N}\sum_{h=1}^{H} \mathbb{E}_{\rho_h}|\widehat{u}_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h)) - u_h^{(i)}(s_h, \widehat{\nu}^{(i)}(s_h))| \\ &\leq \sum_{i=1}^{N}\sum_{h=1}^{H} \mathbb{E}_{\rho_h}\sqrt{|\widehat{u}_h^{(i)}(s_h, \widehat{\nu}_h^{(i)}(s_h)) - u_h^{(i)}(s_h, \widehat{\nu}^{(i)}(s_h))|^2} \\ &\leq \mathcal{O}(HN\sqrt{\log(\mathcal{N}(1/K^2, \mathcal{U}, \|\cdot\|_\infty)\cdot NH/\delta)/K}),\end{aligned}$$

where the second inequality relies on the Cauchy-Schwarz inequality and the last inequality originates from Theorem C.6. Hence we conclude the proof for Theorem 4.4. □

H. Missing Proofs of Auxillary Lemmas

H.1. Proofs for Lemma G.1

Proof of Lemma G.1. Denote one of the ϵ-covering of $\mathcal{F}$ as $\mathcal{F}_\epsilon = \{f_i\}_{i\in[M]} \subset \mathcal{F}$, where $M = \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$. Then applying Bernstein inequality with union bound on the $\mathcal{F}_\epsilon$, it holds with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}[g(X)]\right| \leq \sqrt{\frac{2\mathbb{V}[g(X)]\log(M/\delta)}{n}} + \frac{2R\log(M/\delta)}{3n}, \tag{67}$$

for all $g \in \mathcal{F}_\epsilon$. By the definition of covering number, for any $f \in \mathcal{F}$, there exists $g \in \mathcal{F}_\epsilon$ such that $\|f-g\|_\infty \leq \epsilon$. It then yields that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-g(X_i))\right| + \left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}[g(X)]\right| + |\mathbb{E}[g(X)] - \mathbb{E}[f(X)]| \\ &\leq 2\epsilon + \sqrt{\frac{2\mathbb{V}[g(X)]\log(M/\delta)}{n}} + \frac{2R\log(M/\delta)}{3n}.\end{aligned} \tag{68}$$

Notice that

$$\begin{aligned}\left|\sqrt{\frac{2\mathbb{V}[f(X)]\log(M/\delta)}{n}} - \sqrt{\frac{2\mathbb{V}[g(X)]\log(M/\delta)}{n}}\right| &\leq \sqrt{\left|\frac{2\mathbb{V}[f(X)]\log(M/\delta)}{n} - \frac{2\mathbb{V}[g(X)]\log(M/\delta)}{n}\right|} \\ &= \sqrt{\frac{2\log(M/\delta)}{n}}\cdot\sqrt{|\mathbb{V}[f(X)] - \mathbb{V}[g(X)]|},\end{aligned} \tag{69}$$

where the first inequality is based on the basic inequality $|\sqrt{x}-\sqrt{y}| \leq \sqrt{|x-y|}$ for two absolute variables x, y. What remains is to upper bound the difference of variance in (69).

$$\begin{aligned}|\mathbb{V}[f(X)] - \mathbb{V}[g(X)]| &= \left|\left(\mathbb{E}[(f(X))^2] - (\mathbb{E}[f(X)])^2\right) - \left(\mathbb{E}[(g(X))^2] - (\mathbb{E}[g(X)])^2\right)\right| \\ &= \left|\mathbb{E}\left[(f(X) - \mathbb{E}[g(X)])^2 - (g(X) - \mathbb{E}[f(X)])^2\right]\right| \\ &\leq \mathbb{E}[|f(X) - \mathbb{E}[g(X)] - g(X) + \mathbb{E}[f(X)]|\cdot|f(X) - \mathbb{E}[g(X)] + g(X) - \mathbb{E}[f(X)]|] \\ &\leq 2\epsilon\cdot 4R = 8R\epsilon.\end{aligned} \tag{70}$$

Plugging (69) and (69) into (68), it holds that for all $f \in \mathcal{F}$ with probability at least $1-\delta$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right| &\le 2\epsilon + \sqrt{\frac{2\mathbb{V}[f(X)]\log(M/\delta)}{n}} + \sqrt{\frac{2\log(M/\delta)}{n}}\cdot\sqrt{8R\epsilon} + \frac{2R\log(M/\delta)}{3n} \\ &= 2\epsilon + \sqrt{\frac{2\mathbb{V}[f(X)]\log(M/\delta)}{n}} + 4\sqrt{\frac{R\epsilon\log(M/\delta)}{n}} + \frac{2R\log(M/\delta)}{3n},\end{aligned}$$

which concludes the proof of Lemma G.1. □

H.2. Proofs for Lemma G.2

Proof of Lemma G.2. The proof is adapted from Lemma A.3 in Xie et al. (2021). We first apply Lemma G.1 with $\epsilon = 1/K$ over function class $\mathcal{U}_h^{(i)} = \{(u - u_h^{(i)})^2 : u \in \mathcal{U}\}$ to obtain that

$$\begin{aligned}&\left|\left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}^2 - \left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}^2\right| \\ &\quad= \left|\mathbb{E}_{\rho_h}|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})|^2 - \frac{1}{K}\sum_{\tau=1}^{K}(u(s_h^\tau, x_h^{\tau,(i)}) - u_h^{(i)}(s_h^\tau, x_h^{\tau,(i)}))^2\right| \\ &\quad\le \sqrt{\frac{4\mathbb{V}_{\rho_h}\left[\left(u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right)^2\right]\log(\mathcal{N}(1/K,\mathcal{U}_h^{(i)},\|\cdot\|_\infty)NH/\delta)}{K}} + \frac{8\log(\mathcal{N}(1/K,\mathcal{U}_h^{(i)},\|\cdot\|_\infty)NH/\delta)}{3K} \\ &\qquad+ \frac{8\log(\mathcal{N}(1/K,\mathcal{U}_h^{(i)},\|\cdot\|_\infty)NH/\delta)}{K} + \frac{2}{K} \\ &\quad\le \left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}\cdot\sqrt{\frac{16\log(\mathcal{N}(1/K,\mathcal{U}_h^{(i)},\|\cdot\|_\infty)NH/\delta)}{K}} + \frac{38\log(\mathcal{N}(1/K,\mathcal{U}_h^{(i)},\|\cdot\|_\infty)NH/\delta)}{3K} \\ &\quad\le \left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}\sqrt{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}} + \frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K},\end{aligned} \tag{71}$$

where in the first and second inequality we use the fact that $u \le 1$ for all $u \in \mathcal{U}$ and in the last inequality we use the fact that $\mathcal{N}(1/K,\mathcal{U}_h^{(i)},\|\cdot\|_\infty) \le [\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)]^2$. Here we mark $\mathcal{U}_h^{(i)}$ and $\mathcal{U}$ in red to highlight their difference. Now on the one hand, by basic inequality $|a-b|^2 \le |a^2 - b^2|$, we know from inequality (71) that

$$\begin{aligned}&\left|\left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h} - \left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}\right| \\ &\quad\le \sqrt{\left\|u(c,x^{(i)}) - u_h^{(i)}(c,x^{(i)})\right\|_{2,\rho_h}}\cdot\sqrt[4]{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}} + \sqrt{\frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K}}.\end{aligned} \tag{72}$$

On the other hand, by another basic inequality $|a-b| \le |a - b^2/a|$, we know from inequality (71) that

$$\begin{aligned}&\left|\left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h} - \left\|u(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}\right| \\ &\quad\le \sqrt{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}} + \frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K\cdot\left\|u(c,x^{(i)}) - u_h^{(i)}(c,x^{(i)})\right\|_{2,\rho_h}}.\end{aligned} \tag{73}$$

Thus combining (72) and (73) we obtain that

$$\begin{aligned}
&\left|\left\|u(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}-\left\|u(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}\right| \\
\leq&\min\Bigg\{\sqrt{\left\|u(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}}\cdot\sqrt[4]{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}}+\sqrt{\frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K}}, \\
&\sqrt{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}}+\frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K\cdot\left\|u(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}}\Bigg\}.
\end{aligned} \tag{74}$$

Denote $\eta=\|u(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\|_{2,\rho_h}$ and optimize over $\eta>0$ in (74), we obtain that

$$\begin{aligned}
&\left|\left\|u(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}-\left\|u(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}\right| \\
&\quad\leq\max_{\eta>0}\min\Bigg\{\sqrt{\eta}\cdot\sqrt[4]{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}}+\sqrt{\frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K}}, \\
&\quad\sqrt{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}}+\frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K\cdot\eta}\Bigg\} \\
&\quad\leq\min_{\eta>0}\max\Bigg\{\sqrt{\eta}\cdot\sqrt[4]{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}}+\sqrt{\frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K}}, \\
&\quad\sqrt{\frac{32\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}}+\frac{76\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K\cdot\eta}\Bigg\} \\
&\leq\sqrt{\frac{82\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}},
\end{aligned}$$

where we choose $\eta=\sqrt{\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)/K}$. Here we mark $\max_{\eta>0}\min$ and $\leq\min_{\eta>0}\max$ in red to highlight their difference. This finishes the proof of Lemma G.2. □

H.3. Proofs for Lemma G.3

Proof of Lemma G.3. The proof is adapted from Lemma A.4 in Xie et al. (2021). We first note that we can rewrite

$$\begin{aligned}
&\left\|\widehat{u}(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}^2-\left\|\widetilde{u}(s,x^{(i)})-u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}^2 \\
=&\frac{1}{K}\sum_{\tau=1}^{K}\left(\widehat{u}(s_h^\tau,x_h^{\tau,(i)})-\widetilde{u}(s_h^\tau,x_h^{\tau,(i)})\right)\left(\widehat{u}(s_h^\tau,x_h^{\tau,(i)})+\widetilde{u}(s_h^\tau,x_h^{\tau,(i)})-2u_h^{(i)}(s_h^\tau,x_h^{\tau,(i)})\right).
\end{aligned} \tag{75}$$

By (75), we apply Lemma G.1 with $\epsilon = 1/K$ and function class $\mathcal{U}_h^{\dagger,(i)} = \{(\widehat{u}-\widetilde{u})(\widehat{u}+\widetilde{u}-2u_h^{(i)}) : \widehat{u},\widetilde{u} \in \mathcal{U}\}$ to obtain that

$$
\begin{aligned}
&\Bigg| \left\|\widehat{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}^2 - \left\|\widetilde{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\rho_h}^2 \\
&\quad - \left\|\widehat{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}^2 + \left\|\widetilde{u}(s,x^{(i)}) - u_h^{(i)}(s,x^{(i)})\right\|_{2,\mathcal{D}_h}^2 \Bigg| \\
\leq &\sqrt{\frac{4\mathbb{V}_{\rho_h}\left[\left(\widehat{u}(s,x^{(i)}) - \widetilde{u}(s,x^{(i)})\right)\left(\widehat{u}(s,x^{(i)}) + \widetilde{u}(s,x^{(i)}) - 2u_h^{(i)}(s,x^{(i)})\right)\right]\log(\mathcal{N}(1/K,\mathcal{U}_h^{\dagger,(i)},\|\cdot\|_\infty)NH/\delta)}{K}} \\
&\quad + \frac{16\log(\mathcal{N}(1/K,\mathcal{U}_h^{\dagger,(i)},\|\cdot\|_\infty)NH/\delta)}{3K} + \frac{16\log(\mathcal{N}(1/K,\mathcal{U}_h^{\dagger,(i)},\|\cdot\|_\infty)NH/\delta)}{K} + \frac{2}{K}, \\
\leq &\sqrt{\frac{16\mathbb{V}_{\rho_h}\left[\left(\widehat{u}(s,x^{(i)}) - \widetilde{u}(s,x^{(i)})\right)\left(\widehat{u}(s,x^{(i)}) + \widetilde{u}(s,x^{(i)}) - 2u_h^{(i)}(s,x^{(i)})\right)\right]\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}} \\
&\quad + \frac{64\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K} + \frac{64\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K} + \frac{2}{K}, \\
\leq &\left\|\widehat{u}(s,x^{(i)}) - \widetilde{u}(s,x^{(i)})\right\|_{2,\rho_h} \cdot \sqrt{\frac{64\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{K}} + \frac{262\log(\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)NH/\delta)}{3K},
\end{aligned}
$$

where in the first and second inequality we use the fact that $u \leq 1$ for all $u \in \mathcal{U}$ and the fact that $\mathcal{N}(1/K,\mathcal{U}_h^{\dagger,(i)},\|\cdot\|_\infty) \leq [\mathcal{N}(1/K,\mathcal{U},\|\cdot\|_\infty)]^4$. Here we mark $\mathcal{U}_h^{\dagger,(i)}$ and $\mathcal{U}$ in red to highlight their difference. This finishes the proof of Lemma G.3. □

Proof of Lemma G.7. Before we proceed, we need introduce some concepts to help characterize the convergence rate of MLE estimator, which follows from Geer et al. (2000); Uehara & Sun (2021).

We define the modified function class of $\mathcal{P}_h$:

$$
\overline{\mathcal{P}}_h = \left\{\sqrt{\frac{\widehat{P}+P_h}{2}} \,\middle|\, \widehat{P} \in \mathcal{P}\right\}.
$$

Given a function class $\mathcal{F}$, let $\mathcal{N}_{[]}(\delta,\mathcal{F},\|\cdot\|_{2,\rho_h})$ be the bracketing number of $\mathcal{F}$ w.r.t the norm $\|\cdot\|_{2,\rho_h}$ given by

$$
\|f\|_{2,\rho_h} = \mathbb{E}_{\rho_h}\left[\int \left(f\left(s' \mid s,b\right)\right)^2 \mathrm{d}s'\right]^{1/2}.
$$

Then, the entropy integral of $\mathcal{F}$ is given by

$$
J_B(\delta,\mathcal{F},\|\cdot\|_{2,\rho_h}) = \max\left\{\int_{\delta^2/2}^{\delta}\left(\sqrt{\log \mathcal{N}_{[]}(u,\mathcal{F},\|\cdot\|_{2,\rho_h})}\right)\mathrm{d}u, \delta\right\}.
$$

We also define the localized class of $\overline{\mathcal{P}}_h$:

$$
\overline{\mathcal{P}}_h(\delta) = \left\{\widehat{P} \in \overline{\mathcal{P}}_h : \mathbb{E}_{\rho_h}\left[h^2\left(\widehat{P}(\cdot \mid s,b) \| P_h(\cdot \mid s,b)\right)\right] \leq \delta^2\right\},
$$

where $h\left(\widehat{P}(\cdot \mid s,b) \| P_h(\cdot \mid s,b)\right)$ denotes Hellinger distance defined by

$$
\sqrt{0.5\int\left(\sqrt{\widehat{P}\left(s' \mid s,b\right)} - \sqrt{P_h\left(s' \mid s,b\right)}\right)^2 \mathrm{d}s'}.
$$

Then we introduce the following lemma (Theorem 4 in Uehara & Sun (2021)) to characterize the property of MLE estimator.

Lemma H.1 (MLE guarantee with general function approximation)**.** *We take a function $G_h(\epsilon):[0,1]\to\mathbb{R}$ s.t. $G_h(\epsilon)\geq J_B[\epsilon,\overline{\mathcal{P}}_h(\epsilon),\|\cdot\|_{2,\rho_h}]$ and $G_h(\epsilon)/\epsilon^2$ is a non-increasing function w.r.t ϵ. Then, letting ζ_h be a solution to $\sqrt{K}\epsilon^2\geq c_0 G_h(\epsilon)$ w.r.t ϵ, where c_0 is an absolute constant. With probability at least $1-\delta$, we have*

$$\mathbb{E}_{\rho_h}\left[\left\|\widehat{P}_h^{\mathrm{MLE}}(\cdot\,|\,s,b)-P_h(\cdot\,|\,s,b)\right\|_1^2\right]\leq c_1\left\{\zeta_h+\sqrt{\log\left(c_2/\delta\right)/K}\right\}^2.$$

Proof. Please refer to Theorem 4 in Uehara & Sun (2021). □

Our next step is to show that selecting $\zeta_h=c_2\sqrt{\log\mathcal{N}_{[]}(1/K,\mathcal{P},\|\cdot\|_{1,\infty})/K}$ in Lemma H.1 suffices to prove Lemma G.7. First we show the following facts to discuss the relationship of bracketing numbers of different function classes.

Lemma H.2. *It holds that for all $\epsilon\geq 0$, $\mathcal{N}_{[]}(\epsilon,\overline{\mathcal{P}}_h(\epsilon),\|\cdot\|_{2,\rho_h})\leq\mathcal{N}_{[]}(2\epsilon^2,\mathcal{P},\|\cdot\|_{1,\infty})$.*

Proof. Noticing $\mathcal{N}_{[]}(\epsilon,\overline{\mathcal{P}}_h(\epsilon),\|\cdot\|_{2,\rho_h})\leq\mathcal{N}_{[]}(\epsilon,\overline{\mathcal{P}}_h,\|\cdot\|_{2,\rho_h})$, it suffices to prove that

$$\mathcal{N}_{[]}(\epsilon,\overline{\mathcal{P}}_h,\|\cdot\|_{2,\rho_h})\leq\mathcal{N}_{[]}(2\epsilon^2,\mathcal{P},\|\cdot\|_{1,\infty}).$$

Take the $4\epsilon^2$-brackets of $\mathcal{P}$ as $\mathcal{B}_{\mathcal{P}}=\{(P_j^{\mathrm{U}},P_j^{\mathrm{L}})\}_{j\in[M]}$, where $M=\mathcal{N}_{[]}(4\epsilon^2,\mathcal{P},\|\cdot\|_{1,\infty})$. Then for any $\widetilde{P}_0\in\overline{\mathcal{P}}=\sqrt{\frac{P_0+P_h}{2}}$, there exists $j\in[M]$, s.t. $P_j^{\mathrm{L}}\leq P_0\leq P_j^{\mathrm{U}}$ and $\|P_j^{\mathrm{L}}-P_j^{\mathrm{U}}\|_{1,\infty}\leq 4\epsilon^2$. Hence, $\sqrt{\frac{P_j^{\mathrm{L}}+P_h}{2}}\leq\widetilde{P}_0\leq\sqrt{\frac{P_j^{\mathrm{U}}+P_h}{2}}$. It also holds that,

$$\begin{aligned}\left\|\sqrt{\frac{P_j^{\mathrm{L}}+P_h}{2}}-\sqrt{\frac{P_j^{\mathrm{U}}+P_h}{2}}\right\|_{2,\rho_h}&=\mathbb{E}_{\rho_h}\left[\int_{\mathcal{S}}\left(\sqrt{\frac{P_j^{\mathrm{L}}+P_h}{2}}-\sqrt{\frac{P_j^{\mathrm{U}}+P_h}{2}}\right)^2\mathrm{d}s'\right]^{1/2}\\&\leq\mathbb{E}_{\rho_h}\left[\int_{\mathcal{S}}\left|\frac{P_j^{\mathrm{L}}+P_h}{2}-\frac{P_j^{\mathrm{U}}+P_h}{2}\right|\mathrm{d}s'\right]^{1/2}\\&\leq\sqrt{\frac{1}{2}\left\|P_j^{\mathrm{U}}-P_j^{\mathrm{L}}\right\|_{1,\infty}}\leq\epsilon,\end{aligned}$$

where the first inequality relies on the basic inequality $|\sqrt{a}-\sqrt{b}|\leq\sqrt{|a-b|}$.
Hence, $\left\{\left(\sqrt{\frac{P_j^{\mathrm{L}}+P_h}{2}},\sqrt{\frac{P_j^{\mathrm{U}}+P_h}{2}}\right)\right\}_{j\in[M]}$ are also the ϵ-brackets of $\overline{\mathcal{P}}_h$, which concludes the proof of Lemma H.2. □

In Lemma H.2, we choose $G_h(\epsilon)=(\epsilon-\epsilon^2/2)\sqrt{\log\mathcal{N}_{[]}(\epsilon^4/2,\mathcal{P},\|\cdot\|_{1,\infty})}$, which satisfies that (because of Lemma H.2)

$$\begin{aligned}G_h(\epsilon)&\geq(\epsilon-\epsilon^2/2)\sqrt{\log\mathcal{N}_{[]}(\epsilon^2/2,\overline{\mathcal{P}}_h(\epsilon),\|\cdot\|_{2,\rho_h})}\\&\geq J_B(\epsilon,\overline{\mathcal{P}}_h(\epsilon),\|\cdot\|_{2,\rho_h}),\end{aligned}\tag{76}$$

when we assume that $\log\mathcal{N}_{[]}(\epsilon^2/2,\overline{\mathcal{P}}_h(\epsilon),d)\geq 2$.

It is easy to find that $G(\epsilon)/\epsilon^2$ is non-increasing function. Assuming that $K\geqslant\log\mathcal{N}_{[]}(\epsilon^2/16,\mathcal{P},\|\cdot\|_{1,\infty})$ and solving $\sqrt{K}\epsilon^2\geq c_0G_h(\epsilon)$, we derive the feasible solution region

$$\left\{\epsilon\in[0,1]:\epsilon\geq\frac{c_0}{\sqrt{K}-c_0/2\sqrt{\log\mathcal{N}_{[]}(\epsilon^4/2,\mathcal{P},\|\cdot\|_{1,\infty})}}\right\}.$$

Then there exists an absolute constant c_2, s.t. $\zeta_h=c_2\sqrt{\log\mathcal{N}_{[]}(1/K^2,\mathcal{P},\|\cdot\|_{1,\infty})/K}$ falls into such a feasible region. Hence, by Lemma H.1, there exists a constant c', s.t.

$$\mathbb{E}_{\rho_h}\left[\left\|\widehat{P}_h^{\mathrm{MLE}}(\cdot\,|\,s,b)-P_h(\cdot\,|\,s,b)\right\|_1^2\right]\leq c'\log(\mathcal{N}_{[]}(1/K^2,\mathcal{P},\|\cdot\|_{1,\infty})/\delta)/K.$$

Taking a union bound for $h \in [H]$ and rescaling δ, we obtain that

$$\sup_{h\in[H]} \mathbb{E}_{\rho_h}\left[\left\|\widehat{P}_h^{\mathrm{MLE}}(\cdot\,|\,s,b) - P_h(\cdot\,|\,s,b)\right\|_1^2\right] \le c'\log(\mathcal{N}_{[]}(1/K^2,\mathcal{P},\|\cdot\|_{1,\infty})H/\delta)/K,$$

which concludes the proof of Lemma H.3. □

Proof of Lemma G.8. Motivated by Uehara & Sun (2021), we need to consider the localized class and apply Bernstein inequality to sharpen the convergence rate. We define the estimator localized class as

$$\mathcal{P}_h^{\mathrm{Loc}\,1} := \left\{\widehat{P}\in\mathcal{P} : \mathbb{E}_{\rho_h} g_h(\widehat{P}) \le c'\log(\mathcal{N}_{[]}(1/K^2,\mathcal{P},\|\cdot\|_{1,\infty})H/\delta)/K\right\}. \tag{77}$$

Then we define the corresponding function class

$$\mathcal{F}_h^1 := \{\|\widehat{P}(\cdot\,|\,s,b) - P_h(\cdot\,|\,s,b)\|_1^2 : \widehat{P}\in\mathcal{P}_h^{\mathrm{Loc1}}\}. \tag{78}$$

We denote by $M_1(\epsilon) := \mathcal{N}(\epsilon,\mathcal{F}_h^1,\|\cdot\|_{1,\infty})$ and notice that $\widehat{P}_h^{\mathrm{MLE}} \in \mathcal{P}_h^{\mathrm{Loc}\,1}$ for all $h\in[H]$ on the event E_2 defined in Lemma G.7. Applying Lemma G.1 on the function class $\mathcal{F}_h^1$ with the union bound over $h\in[H]$, it holds for all $h\in[H]$ and $\widehat{P}\in\mathcal{P}_h^{\mathrm{Loc}\,1}$ with probability at least $1-\delta/16$ that

$$\begin{aligned}
\left|(\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h})[g_h(\widehat{P})]\right| &\le 2\epsilon + \sqrt{\frac{2\mathbb{V}_{\rho_h}[g_h(\widehat{P})]\log(M_1(\epsilon)H/\delta)}{K}} + 8\sqrt{\frac{\epsilon\log(M_1(\epsilon)/\delta)}{n}} + \frac{8\log(M_1(\epsilon)/\delta)}{3K} \\
&\le 2\epsilon + \sqrt{\frac{8\mathbb{E}_{\rho_h}[g_h(\widehat{P})]\log(M_1(\epsilon)H/\delta)}{K}} + 8\sqrt{\frac{\epsilon\log(M_1(\epsilon)H/\delta)}{K}} + \frac{8\log(M_1(\epsilon)H/\delta)}{3K} \\
&\le 2\epsilon + \frac{\sqrt{8c'\log(\mathcal{N}_{[]}(1/K,\mathcal{P},\|\cdot\|_{1,\infty})H/\delta)\cdot\log(M_1(\epsilon)H/\delta)}}{K} \\
&\quad + 8\sqrt{\frac{\epsilon\log(M_1(\epsilon)H/\delta)}{K}} + \frac{8\log(M_1(\epsilon)H/\delta)}{3K},
\end{aligned} \tag{79}$$

where the first inequality also relies on the fact that $\sup_{\widehat{P}\in\mathcal{P}}\|g_h(P)\|_\infty \le \sup_{\widehat{P}\in\mathcal{P}}(\|\widehat{P}\|_{1,\infty} + \|P_h\|_{1,\infty})^2 \le 4$. To select a proper ϵ, we define a larger function class $\mathcal{F}_h^0$ as follows,.

$$\mathcal{F}_h^0 := \{\|\widehat{P}(\cdot\,|\,s,b) - P_h(\cdot\,|\,s,b)\|_1^2 : \widehat{P}\in\mathcal{P}\}. \tag{80}$$

By the following lemma, we characterize the relationship of $\mathcal{F}_h^0$ and $\mathcal{F}_h^1$.

Lemma H.3. *It holds for all $h\in[H]$ that $\mathcal{N}(\epsilon,\mathcal{F}_h^0,\|\cdot\|_\infty) \le \mathcal{N}_{[]}(\epsilon,\mathcal{P},\|\cdot\|_{1,\infty})$.*

Proof. For any $\widehat{P}\in\mathcal{P}$, there exists $P_i^{\mathrm{Cover}} \in \{P_j^{\mathrm{Cover}}\}_{j\in[M]} \subset \mathcal{P}$, where $M = \mathcal{N}(\epsilon,\mathcal{P},\|\cdot\|_{1,\infty})$, s.t. $\|P_i^{\mathrm{Cover}} - \widehat{P}\|_{1,\infty} \le \epsilon$. Notice that

$$\begin{aligned}
\left|\left[\|P_i^{\mathrm{Cover}} - P_h\|_1^2 - \|\widehat{P} - P_h\|_1^2\right](s,b)\right| &\le 2\left|\left[\|P_i^{\mathrm{Cover}} - P_h\|_1 - |\|\widehat{P} - P_h\|_1\right](s,b)\right| \\
&\le 2\left|\left[\|P_i^{\mathrm{Cover}} - \widehat{P}\|_1\right](s,b)\right| \\
&\le 2\|P_i^{\mathrm{Cover}} - \widehat{P}\|_{1,\infty} \\
&\le 2\|P_i^{\mathrm{Cover}} - \widehat{P}\|_{1,\infty} \le 2\epsilon.
\end{aligned} \tag{81}$$

Taking supreme over $\mathcal{S}\times\mathcal{B}$, we obtain that $\left\|\left[\|P_i^{\mathrm{Cover}} - P_h\|_1^2 - \|\widehat{P} - P_h\|_1^2\right]\right\|_\infty \le 2\epsilon$, which implies that

$$\mathcal{N}(2\epsilon,\mathcal{F}_h^0,\|\cdot\|_\infty) \le \mathcal{N}(\epsilon,\mathcal{P},\|\cdot\|_{1,\infty}).$$

Notice that covering number can be upper bounded by bracketing number, that is,

$$\mathcal{N}(2\epsilon,\mathcal{F}_h^0,\|\cdot\|_\infty) \le \mathcal{N}(\epsilon,\mathcal{P},\|\cdot\|_{1,\infty}) \le \mathcal{N}_{[]}(2\epsilon,\mathcal{P},\|\cdot\|_{1,\infty}),$$

which concludes the result of Lemma H.3. □

Since $M_1(\epsilon) \leq \mathcal{N}(\epsilon, \mathcal{F}_h^0, \|\cdot\|_\infty) \leq \mathcal{N}_{[]}(\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty})$, selecting a proper $\epsilon = 1/K^2$, we have with probability at least $1-\delta/16$ that

$$\sup_{h\in[H]} \sup_{\widehat{P}\in\mathcal{P}_h^{\text{Loc}\,1}} |[\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h}] g_h(\widehat{P})| \leq c'' \log(\mathcal{N}_{[]}(1/K^2, \mathcal{P}, \|\cdot\|_{1,\infty}) H/\delta)/K,$$

where c'' is an absolute constant. Hence we finish the proof of Lemma G.8. □

Proof of Lemma G.9. This proof is more complicated than the proof of Lemma G.8. On the event E_1 defined in Lemma G.6, we define the estimator localized class as

$$\mathcal{P}_h^{\text{Loc}\,2} := \{\widehat{P} \in \mathcal{P}_{h,\xi_2} : \mathbb{E}_{\mathcal{D}} g_h(\widehat{P}) \leq \xi_2\}. \tag{82}$$

We also define the function class

$$\mathcal{F}_h^2 := \{\|\widehat{P}(\cdot\,|\,s,b) - P_h(\cdot\,|\,s,b)\|_1^2 : \widehat{P} \in \mathcal{P}_h^{\text{Loc}2}, \text{ for all } h \in [H]\}. \tag{83}$$

We denote by $M_2(\epsilon) := \mathcal{N}(\epsilon, \mathcal{F}_h^2, \|\cdot\|_{1,\infty})$ and notice that $\widehat{P}_h \in \mathcal{P}_h^{\text{Loc}\,2}$ on the event E_1 defined in Lemma G.6. Applying Lemma G.1 on $\mathcal{F}_h^2$ with union bound over $h \in [H]$, we have for all $h \in [H]$ and $\widehat{P} \in \mathcal{P}_h^{\text{Loc}\,2}$ with probability at least $1-\delta/16$ that

$$\begin{aligned}
\big|(\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h})[g_h(\widehat{P})]\big| &\leq 2\epsilon + \sqrt{\frac{2\mathbb{V}_{\rho_h}[g_h(\widehat{P})]\log(M_2(\epsilon)H/\delta)}{K}} + 8\sqrt{\frac{\epsilon\log(M_2(\epsilon)/\delta)}{n}} + \frac{8\log(M_2(\epsilon)/\delta)}{3K} \\
&\leq 2\epsilon + \sqrt{\frac{8\mathbb{E}_{\rho_h}[g_h(\widehat{P})]\log(M_2(\epsilon)H/\delta)}{K}} + 8\sqrt{\frac{\epsilon\log(M_2(\epsilon)H/\delta)}{K}} + \frac{8\log(M_2(\epsilon)H/\delta)}{3K} \\
&\leq 2\epsilon + \sqrt{\frac{8(|[\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h}]g_h(\widehat{P})| + \xi_2)}{K}\log(M_2(\epsilon)H/\delta)} \\
&\quad + 8\sqrt{\frac{\epsilon\log(M_2(\epsilon)H/\delta)}{K}} + \frac{8\log(M_2(\epsilon)H/\delta)}{3K}.
\end{aligned} \tag{84}$$

By Lemma H.3, it holds that $M_2(\epsilon) \leq \mathcal{N}(\epsilon, \mathcal{F}_h^1, \|\cdot\|_\infty) \leq \mathcal{N}_{[]}(\epsilon, \mathcal{P}, \|\cdot\|_{1,\infty})$. Selecting a proper $\epsilon = 1/K^2$, we solve the quadratic inequality (84) with respect to $|[\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h}]g_h(\widehat{P})|$. We obtain that (Uehara & Sun, 2021; Xie et al., 2021)

$$\sup_{h\in[H]} \sup_{\widehat{P}\in\mathcal{P}_h^{\text{Loc}\,2}} |[\mathbb{E}_{\mathcal{D}_h} - \mathbb{E}_{\rho_h}] g_h(\widehat{P})| \leq C'' \log(\mathcal{N}_{[]}(1/K^2, \mathcal{P}, \|\cdot\|_{1,\infty}) H/\delta)/K$$

with probability at least $1-\delta/16$, where C'' is an absolute constant. Hence we conclude the proof of Lemma G.9. □

I. Useful Lemmas for Reproducing Kernel Hilbert Space

Lemma I.1 (Covering Number of RKHS Ball under $\|\cdot\|_\infty$-Norm)**.** *Under Assumption D.9, the covering number of RKHS ball $\mathcal{H}_R = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$ with radious R under $\|\cdot\|_\infty$-norm is bounded by*

$$\log \mathcal{N}(\epsilon, \mathcal{H}_R, \|\cdot\|_\infty) \leq C \cdot \log^2(1/\gamma)/\gamma \cdot \log^{1+1/\gamma}(R/\epsilon).$$

where $C > 0$ is an absolute constant.

Proof of Lemma I.1. See Lemma C.2. in Cai et al. (2020b) for a detailed proof. □

Lemma I.2 (Eluder Dimension: RKHS)**.** *Under Assumption D.9, the eluder dimension of function class $\mathcal{F}$ with functions upper bounded by M parameterized by RKHS ball $\mathcal{H}_R$ with radious R can be bounded by*

$$\dim_{\mathbb{E}}(\mathcal{F}, \epsilon) \leq C \cdot \log^2(1/\gamma)/\gamma \cdot \log^{1+1/\gamma}(RM/\epsilon).$$

Proof of Lemma I.2. See Lemma C.1. in Cai et al. (2020b) for a detailed proof. □

Lemma I.3 (RKHS Trancation Error with Assumption D.9)**.** *Let C_1 and C_2 be the absolute constants in Assumption D.9. There exists an absolute constant $\widetilde{C}$ such that for any $\gamma \in (0, 1/2), t \geq 1$, and $R \geq 2$, if we set*

$$d_0 = \left\lceil \widetilde{C} \cdot \log(1/\gamma)/\gamma \cdot \log^{1/\gamma}(tR) \right\rceil,$$

then it holds that $d_0^\gamma \geq 4(1-\gamma)\left(\gamma C_2\right)^{-1}$ and

$$\varepsilon_{d_0} := \sum_{j > d_0} \sqrt{\lambda_j} \cdot R \leq C_1^{1/2} d_0^{1-\gamma} R \left(\gamma C_2\right)^{-1} \cdot \exp\left(-C_2 d_0^\gamma/2\right) \leq 1/t.$$

Proof of Lemma I.3. See Lemma F.7. in Cai et al. (2020b) for a detailed proof. □