A Self-Play Posterior Sampling Algorithm for Zero-Sum Markov Games

Wei Xiong [1] **Han Zhong** [2] **Chengshuai Shi** [3] **Cong Shen** [3] **Tong Zhang** [1 4]

Abstract

Existing studies on provably efficient algorithms for Markov games (MGs) almost exclusively build on the "optimism in the face of uncertainty" (OFU) principle. This work focuses on a different approach of posterior sampling, which is celebrated in many bandits and reinforcement learning settings but remains under-explored for MGs. Specifically, for episodic two-player zero-sum MGs, a novel posterior sampling algorithm is developed with *general* function approximation. Theoretical analysis demonstrates that the posterior sampling algorithm admits a $\sqrt{T}$-regret bound for problems with a low multi-agent decoupling coefficient, which is a new complexity measure for MGs, where T denotes the number of episodes. When specialized to linear MGs, the obtained regret bound matches the state-of-the-art results. To the best of our knowledge, this is the first provably efficient posterior sampling algorithm for MGs with frequentist regret guarantees, which enriches the toolbox for MGs and promotes the broad applicability of posterior sampling.

1. Introduction

Multi-agent reinforcement learning (MARL) focuses on the sequential decision-making problem involving more than one agent, each of which aims to optimize her own long-term return by interacting with the environment and other agents (Zhang et al., 2021). Today, MARL has a diverse set of real-world applications, including Go (Silver et al., 2016; 2017), autonomous driving (Shalev-Shwartz et al., 2016), Poker (Brown & Sandholm, 2019), and Dota (Berner et al., 2019), just to name a few. Due to the large state space of these practical problems, function approximation (with neural networks) is often used in these applications for the generalization across different state-action pairs. While there is a long line of related works on the theoretical understanding of single-agent RL with general function approximation (Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020; Jin et al., 2021a; Du et al., 2021; Dann et al., 2021), the theory of MARL with general function approximation is substantially less explored. In this paper, we aim to explore this topic in the context of two-player zero-sum Markov games (MGs) (Shapley, 1953; Littman, 1994).

The goal of learning in a two-player zero-sum MG is to learn the Nash equilibrium at which the policy of each player maximizes her own cumulative rewards, provided that the policies of other agents are fixed. Intuitively speaking, Nash equilibrium characterizes the point from which no agent will deviate. Since the reward and the state transition are determined jointly by the actions of both agents, in addition to the unknown environment, each agent must also handle the dynamics of other strategic agents. Due to this game-theoretical feature, algorithms designed for MDP cannot be directly extended to the MARL case. However, recent studies (Jin et al., 2021b; Huang et al., 2021) have shown that with an innovative asymmetrical structure, similar theoretical results can be established for the two-player zero-sum MG with general function approximation.

Nevertheless, despite a handful of recent progress on the theory of the two-player zero-sum MG with general function approximation, the existing works are mainly confined to algorithms based on the optimism in the face of uncertainty (OFU) principle. In contrast, the theory of posterior-sampling-based algorithms is less developed (in the frequentist setting). Various empirical studies indicate that the OFU-based algorithms can be far too optimistic for average instances and is inferior to posterior sampling algorithms, including Chapelle & Li (2011) for bandit, and Osband et al. (2016) for RL. Recent works in the context of contextual multi-armed bandit and single-agent RL demonstrate that there is no statistical efficiency gap between OFU and posterior sampling algorithms (Dann et al., 2021; Zhang, 2021). However, whether we can design model-free posterior sampling algorithms in MARL that achieve similar theoretical guarantees remains open.

In this paper, we are interested in the application of posterior sampling in the two-player zero-sum MG with general

[1]The Hong Kong University of Science and Technology; [2]Center for Data Science, Peking University; [3]University of Virginia; [4]Google Research. Correspondence to: Tong Zhang <tongzhang@tongzhang-ml.org>.

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

function approximation and self-play (which means that the learning agent can control both players). Our main result indicates that, similar to the single-agent case, posterior sampling algorithms can achieve comparable theoretical guarantees as to the OFU-based algorithms. Our contributions are summarized as follows:

- A provably efficient posterior sampling algorithm is designed under the self-play framework for the two-player zero-sum MG with general function approximation. To the best of our knowledge, this is the first posterior sampling algorithm with frequentist regret guarantee in the context of Markov games;
- The single-agent complexity measure of *decoupling coefficient*, first introduced in Dann et al. (2021), is extended to the multi-agent setting, Moreover, a number of examples with provably small multi-agent decoupling coefficients are identified;
- The proposed algorithm is rigorously proved to obtain a $\sqrt{T}$-regret for problems with low multi-agent decoupling coefficient, where T is the number of episodes.

It is noted that the sampling procedure of the proposed algorithm may not be computationally efficient. The lack of computational tractability also appears in the works of Jin et al. (2021b); Huang et al. (2021), as well as previous works with general function approximation in the context of single-agent RL (Jiang et al., 2017; Jin et al., 2021a; Dann et al., 2021; Du et al., 2021). It is an interesting future research topic to identify cases where efficient sampling is possible. Moreover, we do not take credit for the asymmetrical structure in our algorithmic framework. The main contribution here is to extend the posterior sampling algorithm to MGs under the self-play framework.

1.1. Related Works

There have been a lot of works focusing on designing provably efficient algorithms for zero-sum MGs. For the tabular setting, Bai et al. (2020); Bai & Jin (2020); Liu et al. (2020) provide $O(\text{poly}(|\mathcal{X}|, |\mathcal{A}|, |\mathcal{B}|, H) \cdot \sqrt{T})$ regret guarantees for the proposed algorithms, where $|\mathcal{X}|$ is the number of states, $|\mathcal{A}|$ and $|\mathcal{B}|$ are the number of action spaces of two players, respectively, H is the episode length, and T is the number of episodes. Then, Xie et al. (2020); Chen et al. (2021) study two linear-type MGs and design algorithms with $O(\text{poly}(d, H)\cdot\sqrt{T})$ regret, where d is the dimension of the linear features. Recently, Jin et al. (2021b); Huang et al. (2021) further propose efficient algorithms for zero-sum MGs with general function approximation.

Our work is also closely related to another line of work on posterior sampling algorithms. In the context of contextual bandit, due to the impressive empirical performance of Thompson Sampling (Chapelle & Li, 2011; Osband et al., 2016), there have been significant efforts in developing its theoretical analysis, including Russo & Van Roy (2014) in the form of Bayesian regret and Kaufmann et al. (2012); Zhang (2021) in the frequentist setting. For the Markov Decision Process (MDP), the seminal work Osband & Van Roy (2014) considers the Bayesian regret and proposes a general posterior sampling RL method. The randomized least-squares value iteration (RLSVI) algorithm (Osband et al., 2016) is shown to admit frequentist regret bounds for tabular MDP (Russo, 2019; Agrawal et al., 2020; Xiong et al., 2021) and linear MDP (Zanette et al., 2020). Beyond the linear setting, a recent work (Dann et al., 2021) proposes a conditional posterior sampling algorithm to solve the MDP with general function approximation.

A recent posterior-sampling-type work by Jafarnia-Jahromi et al. (2021) considers the infinite-horizon zero-sum MGs with average-reward criterion in the tabular setting, with a focus on the Bayesian regret, whose analysis technique is fundamentally different from ours. To the best of our knowledge, there is no posterior sampling algorithm with a frequentist regret guarantee to date for MGs.

2. Problem Formulation

Markov Games (MGs) generalize the standard Markov Decision Processes to the multi-agent setting. In this work, the episodic two-player zero-sum MG is considered, which can be formally denoted as $MG(H, \mathcal{X}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r)$. Here H denotes the length of each episode, $\mathcal{X}$ is the (possibly infinite) state space, $\mathcal{A}$ and $\mathcal{B}$ are the action spaces of two players (referred to as the max-player and the min-player), respectively, $\mathbb{P}_h(\cdot|x, a, b)$ is the transition measure of the next state from the current state x with two actions (a, b) taken at step h, and $r^h(x, a, b)$ is the corresponding reward received with actions (a, b) taken for state x at step h.

Specifically, in this MG, each episode t starts from an initial state x_t^1. At each step h, two players observe the current state x_t^h, take actions (a_t^h, b_t^h) individually, and observe the next state $x_t^{h+1} \sim \mathbb{P}_h(\cdot|x_t^h, a_t^h, b_t^h)$. The current episode ends after step H and then a new episode starts. Without loss of generality, each episode is assumed to have a fixed initial state $x_t^1 = x^1$, which can be easily generalized to having x_t^1 sampled from a fixed but unknown distribution.

Also, for the ease of presentation, the reward $r^h(x, a, b)$ is assumed to be deterministic and in the interval of $[0, 1]$ for any (x, a, b) in this paper, while the algorithm designs and theoretical results can also be applied for stochastic bounded rewards with slight modifications.

Policies and Value Functions. With $\Delta_{\mathcal{A}}$ denoting the probability simplex over the action space $\mathcal{A}$, a Markov policy μ of the max-player can be defined as $\mu := \{\mu_h : \mathcal{X} \to \Delta_{\mathcal{A}}\}_{h\in[H]}$. Similarly, we can define a Markov policy $\nu := \{\nu_h : \mathcal{X} \to \Delta_{\mathcal{B}}\}_{h\in[H]}$ for the min-player.

Given a policy pair (μ, ν), the value function $V_h^{\mu,\nu} : \mathcal{X} \to \mathbb{R}$ at step h is defined as

$$V_h^{\mu,\nu}(x) := \mathbb{E}_{\mu,\nu}\left[\sum_{h'=h}^{H} r^{h'}\left(x^{h'}, a^{h'}, b^{h'}\right) \mid x^h = x\right]$$

and the Q-value function $Q_h^{\mu,\nu} : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ as

$$Q_h^{\mu,\nu}(x,a,b) := \mathbb{E}_{\mu,\nu}\Bigg[\sum_{h=h}^{H} r^{h'}\left(x^{h'}, a^{h'}, b^{h'}\right) \mid (x^h, a^h, b^h) = (x,a,b)\Bigg],$$

where the expectations are taken over the randomness of the environment and the policies.

For a clean presentation, we use the notation $\mathbb{P}_h$ (with a slight abuse) so that $[\mathbb{P}_h V](x,a,b) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a,b)} V(x')$ for any value function V. Similarly, the notation $\mathbb{D}_\pi$ is adopted so that $[\mathbb{D}_\pi Q](x) := \mathbb{E}_{(a,b)\sim\pi(\cdot,\cdot|x)} Q(x,a,b)$, for any policy pair $\pi = (\mu, \nu)$ and action-value function Q. With these notations, the Bellman equations are given by

$$Q_h^{\mu,\nu}(x,a,b) = \left[r^h + \mathbb{P}_h V_{h+1}^{\mu,\nu}\right](x,a,b);$$
$$V_h^{\mu,\nu}(x) = \left[\mathbb{D}_{\mu_h \times \nu_h} Q_h^{\mu,\nu}\right](x).$$

Best Response. For any policy of max-player μ, a corresponding best response for the min-player can be found, denoted as $\nu^\dagger(\mu)$, such that $V_h^{\mu,\nu^\dagger(\mu)}(x) = \inf_\nu V_h^{\mu,\nu}(x)$ for all (x,h). This value is the best favorable result for the min-player if the max-player announces that she will play strategy μ. Similarly, for a min-player policy ν, there exists a best response for the max-player, denoted as $\mu^\dagger(\nu)$, such that $V_h^{\mu^\dagger,\nu}(x) = \sup_\mu V_h^{\mu,\nu}(x)$ for all (x,h). To simplify the notation, we use

$$V_h^{\mu,\dagger}(x) := V_h^{\mu,\nu^\dagger(\mu)}(x), Q_h^{\mu,\dagger}(x,a,b) := Q_h^{\mu,\nu^\dagger(\mu)}(x,a,b);$$
$$V_h^{\dagger,\nu}(x) := V_h^{\mu^\dagger(\nu),\nu}(x), Q_h^{\dagger,\nu}(x,a,b) := Q_h^{\mu^\dagger(\nu),\nu}(x,a,b).$$

Nash Equilibrium. Moreover, there exists a set of Nash equilibrium (NE) policies (μ^*, ν^*) (Filar & Vrieze, 2012) that are optimal against their best response such that

$$V_h^{\mu^*,\dagger}(x) = \sup_\mu V_h^{\mu,\dagger}(x), \quad V_h^{\dagger,\nu^*}(x) = \inf_\nu V_h^{\dagger,\nu}(x),$$

for all $(x,h) \in \mathcal{X} \times [H]$. For this NE, the following famous minimax equation holds:

$$\sup_\mu \inf_\nu V_h^{\mu,\nu}(x) = V_h^{\mu^*,\nu^*}(x) = \inf_\nu \sup_\mu V_h^{\mu,\nu}(x)$$

for all $(x,h) \in \mathcal{X} \times [H]$. For simplicity, we denote $V_h^*(x) := V_h^{\mu^*,\nu^*}(x)$ and $Q_h^*(x) := Q_h^{\mu^*,\nu^*}(x)$. Note that although there might exist multiple NE policies, the NE value function is unique for a zero-sum MG.

Performance metrics. A max-player's policy μ is said to be ϵ-close to the NE if it satisfies $V^*(x^1) - V^{\mu,\dagger}(x^1) < \epsilon$. Note that we have $V^*(x^1) - V^{\mu,\nu}(x^1) \le V^*(x^1) - V^{\mu,\dagger}(x^1)$ for all min-player's policy ν as the best response is the strongest opponent for the max-player. The main goal of this paper is to find an ϵ-close policy for the max-player and her regret over T episodes can be defined as

$$\mathrm{Reg}(T) := \sum_{t=1}^{T}\left[V_1^*(x_1) - V_1^{\mu_t,\dagger}(x_1)\right],$$

where μ_t is the policy adopted by the max-player for episode t. Note that we can switch the roles of two players to learn a policy ν that is ϵ-close to the NE for the min-player.

2.1. Function Approximation

As mentioned in Sec. 1, real-world applications of RL often encounter the challenge of a large state space where storing a table as in the classical Q-learning is generally infeasible. To overcome this challenge, function approximation is proposed and proven to be efficient with many practical successes. Following similar attempts in MDP, we aim to approximate the Q-value functions for the MGs considered in this work by a class of functions $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h \subset (\mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R})$.

For $f \in \mathcal{F}$, a NE can be induced and the corresponding policy μ_f of the max-player is defined for all (x,h) as

$$\mu_{f,h}(x) = \mathrm{argmax}_{\mu\in\Delta_\mathcal{A}} \min_{\nu\in\Delta_\mathcal{B}} \mu^\top f^h(x,\cdot,\cdot)\nu.$$

The induced value function for all (x,h) is then given by

$$V_{f,h}(x) = \max_{\mu\in\Delta_\mathcal{A}} \min_{\nu\in\Delta_\mathcal{B}} \mu^\top f^h(x,\cdot,\cdot)\nu.$$

Moreover, for a fixed max-player policy μ and a function $f \in \mathcal{F}$, the induced value function of the best response of the min-player is defined for all (x,h) as

$$V_{f,h}^\mu(x) = \min_{\nu\in\Delta_\mathcal{B}} \mu_h(x)^\top f^h(x,\cdot,\cdot)\nu.$$

This is mainly for the min-player to choose her policy, given the max-player's policy μ. Details can be found in Sec. 3.

As common in Perolat et al. (2015); Jin et al. (2021b); Huang et al. (2021), two types of Bellman operators are defined as

$$\left(\mathcal{T}_h f\right)(x,a,b) := \left[r^h + \mathbb{P}_h V_{f,h+1}\right](x,a,b);$$
$$\left(\mathcal{T}_h^\mu f\right)(x,a,b) := \left[r^h + \mathbb{P}_h V_{f,h+1}^\mu\right](x,a,b).$$

The corresponding Bellman residual are denoted as

$$\begin{aligned}\mathcal{E}_h(f;x,a,b) &= \mathcal{E}(f^h, f^{h+1}; x, a, b)\\ &= f^h(x,a,b) - (\mathcal{T}_h f)(x,a,b);\\ \mathcal{E}_h^\mu(f;x,a,b) &= \mathcal{E}^\mu(f^h, f^{h+1}; x, a, b)\\ &= f^h(x,a,b) - (\mathcal{T}_h^\mu f)(x,a,b).\end{aligned} \tag{2.1}$$

Sometimes the state-action pair (x,a,b) may be replaced with a trajectory $\zeta = \{(x^{h'}, a^{h'}, b^{h'}, r^{h'})\}_{h'=1}^H$, which indicates that the corresponding state-action pair at step h, i.e. (x^h, a^h, b^h), is taken as input.

Recent advances show that RL with function approximation is, in general, intractable without any further assumption (Krishnamurthy et al., 2016; Weisz et al., 2021). It is thus common to adopt additional assumptions over the function class in the literature on general function approximation in MDPs, especially the realizability and completeness assumptions (Wang et al., 2020; Jin et al., 2021a; Dann et al., 2021). As MGs are natural extensions of MDPs, the generalized realizability and completeness assumptions are also adopted in this work. Note that Assumptions 1 and 2 are also required by other recent works on MGs with general function approximation (Jin et al., 2021b; Huang et al., 2021).

Assumption 1 (Realizability). *For the Nash equilibrium, it holds that $Q_h^* \in \mathcal{F}_h, \forall h \in [H]$. Moreover, for any $f \in \mathcal{F}$, it holds that $Q_h^{\mu_f,\dagger} \in \mathcal{F}_h, \forall h \in [H]$.*

The realizability assumption states that the function class $\mathcal{F}$ is large enough so that it contains the Q-value function of the NE and also the Q-value function of any induced policy and its best response.

The completeness assumption is more restrictive where the main drawback is that completeness is non-monotone, meaning that adding one function into $\mathcal{F}$ may violate the assumption. However, it is the key to handling the variance of sampling in the literature and analysis without completeness seems very challenging.

Assumption 2 (Completeness). *For any $f, g \in \mathcal{F}$ and the induced policy μ_f, it holds that $\mathcal{T}_h^{\mu_f} g \in \mathcal{F}_h, \forall h \in [H]$.*

Additionally, the following boundedness assumption[1] is considered, which is natural for bounded rewards and a finite episode length.

Assumption 3 (Boundedness). *There exists $\beta > 1$ s.t. $f^h(x,a,b) \in [0, \beta-1], \forall (f,h,x,a,b) \in \mathcal{F} \times [H] \times \mathcal{X} \times \mathcal{A} \times \mathcal{B}$.*

3. Algorithm

The proposed *Conditional Posterior Sampling with Booster* algorithm is presented in this section. "Conditional" refers to the design of $q(f^h | f^{h+1}, S_t)$ (the denominator in Eqn. (3.2)), which allows us to use the true Bellman operator $\mathcal{T}_h$ in the analysis even though we do dot know it in the executed algorithm. "Booster" is a synonym for exploiter (Jin et al., 2021b) in the context and refers to the asymmetric structure as the second agent aims to assist the main agent's learning. Another reason is that when we wrote this paper, one of the authors had a fever due to the booster vaccine, and another author tested positive for covid.

Algorithm 1 Conditional Posterior Sampling with Booster

1: **Input**: function class: $\mathcal{F}$, learning rate η, horizon T, prior parameter λ.
2: S_0 is initialized to be empty.
3: **for** Stage $t = 1, \ldots, T$ **do**
4: Main agent: $\mu_t \leftarrow$ Main($\mathcal{F}, \eta, S_{t-1}, T, \lambda$);
5: Booster agent: $\nu_t \leftarrow$ Booster($\mathcal{F}, \eta, S_{t-1}, \mu_t, T, \lambda$);
6: Execute the policies (μ_t, ν_t) and collect the trajectory $(x_t^1, a_t^1, b_t^1, r_t^1, \cdots, x_t^H, a_t^H, b_t^H, r_t^H)$ to obtain S_t.
7: **end for**

3.1. Overview

The existing algorithms with frequentist guarantees are confined to OFU-based algorithms. Algorithmically, these algorithms typically maintain a confidence set $\mathcal{C}$ whose components are empirically consistent with the Bellman equation so far. Then, an optimistic function $f \in \mathcal{C}$ is selected to approximate the true value function through some optimization subroutine (Jin et al., 2021b; Huang et al., 2021). In contrast, the posterior sampling algorithm starts with a *prior* $p_0(\cdot)$ over the function class $\mathcal{F}$ and collects trajectories to compute the *likelihood*; they together lead to a posterior distribution $p(\cdot)$ over $\mathcal{F}$. Then, a function is sampled from the posterior distribution to approximate the target. In addition to the difference in algorithm structure, the analysis techniques for the posterior sampling algorithm are also different, particularly due to the lack of explicit optimism from the planning step.

A frequentist theoretical guarantee of posterior sampling algorithms has been lacking for a long time, even in the context of contextual bandit. Recently, Zhang (2021) and Dann et al. (2021) show that adding an extra optimistic term can lead to frequentistly optimal posterior sampling algorithm in contextual bandit and MDP, respectively. However, in the MARL setting, the multi-agent nature leads to complicated statistical dependence across the players. In particular, in addition to the environment, the agent will also be affected by other strategic agents. Therefore, the situation is more complicated even in the two-player case and the algorithms designed for MDPs cannot be directly extended to MGs. To overcome this issue, inspired by Jin et al. (2021b); Huang et al. (2021), we leverage the innovative asymmetric structure to pick the max-player and the min-player as the main agent and the booster agent, respectively, where the booster

[1] $(\beta - 1)$ is usually assumed to be either 1 or H in the literature.

agent, as the name suggests, aims to assist the main agent's learning.

Our algorithm is summarized in Algorithm 1 where the main agent's algorithm and the booster agent's algorithm are given in Algorithms 2 and 3, respectively.

Algorithm 2 Main($\mathcal{F}, \eta, \mathcal{D}, T, \lambda$)

1: Draw $f \sim p(\cdot|\mathcal{D})$ where the posterior is given by Eqn. (3.3);
2: $\mu_{f,h}(x) = \operatorname{argmax}_{\mu \in \Delta_{\mathcal{A}}} \min_{\nu \in \Delta_{\mathcal{B}}} \mu^\top f^h(x,\cdot,\cdot)\nu, \forall (x,h)$;
3: Return μ_f.

3.2. The Main Agent

The main agent's goal is to learn a ϵ-close policy for the max-player, i.e., $V^*(x^1) - V^{\mu,\dagger}(x^1) < \epsilon$. With function class $\mathcal{F}$ available, she aims to find a function $f \in \mathcal{F}$ to approximate the Nash Q-value function, i.e., Q^*, which can be used to solve the Nash policy via the minimax equation. The following optimistic prior and temporal difference error likelihood are carefully crafted to induce a desired posterior distribution over $\mathcal{F}$, which is further used to sample a suitable function f.

Optimistic prior. The following prior $\tilde{p}_0(\cdot)$ over the function class $\mathcal{F}$ is adopted for the main agent:

$$\tilde{p}_0(f) \propto \exp(\lambda V_{f,1}(x^1)) \prod_{h=1}^{H} p_0^h(f^h), \tag{3.1}$$

where $\lambda > 0$ is a tuning parameter, and $p_0^h(\cdot)$ is a distribution over $\mathcal{F}_h$. Note that other than the standard prior of $p_0(f) = \prod_{h=1}^H p_0^h(f^h)$, an additional *optimistic* term, i.e., $\exp(\lambda V_{f,1}(x^1))$, is involved in the prior, which plays an important role of encouraging exploration for the main agent.

This prior is referred to as an optimistic one because it favors large values for the initial state. Also, technically, it compensates for one extra term arising in the value decomposition in Lemma 1 when the optimism is not inherently available as in OFU-based algorithms. Similar techniques are also adopted in the design of posterior sampling for MDPs (Dann et al., 2021) and contextual bandits (Zhang, 2021). Furthermore, Zhang (2021) argues that in the context of contextual bandit, such an optimistic component is necessary to design optimal posterior-sampling-based algorithms in the frequentist setting.

Also, apart from the optimism itself, the *global* optimism mechanism, meaning that we only add an optimistic term in the prior distributions at the initial value, is the key to achieving improvement in the feature dimension for linear MGs. We will return to this in Sec. 5.3.

Likelihood for the main agent. If we denote the history up to the end of episode t as $S_t = \{x_s^h, a_s^h, b_s^h, r_s^h\}_{s\in[t],h\in[H]}$, a likelihood over S_t is specified as

$$p(S_t|f) \propto \prod_{h=1}^{H} \frac{\exp\left(-L^h(f^h, f^{h+1}; S_t)\right)}{\mathbb{E}_{f^h\sim p_0^h} \exp(-\eta L^h(f^h, f^{h+1}; S_t))}. \tag{3.2}$$

$\{L^h(\cdot)\}_{h=1}^H$ is a collection of squared loss functions as

$$L^h(f^h, f^{h+1}; S_t) = \sum_{s=1}^{t} \left[f^h(x_s^h, a_s^h, b_s^h) - r_s^h - V_{f^{h+1}}(x_s^{h+1})\right]^2,$$

which is a proxy to the squared $\mathcal{T}_h$-Bellman error. The likelihood in Eqn. (3.2) introduces a special denominator, which is motivated by that for MDP (Dann et al., 2021). We will see that the denominator is the key to handling the variance of sampling, but that is also why we need the strong completeness assumption. We will discuss this in Sec. 4.3.

Posterior distribution for the main agent. Given the prior distribution and the likelihood, the posterior at the end of episode t can be naturally expressed as

$$p(f|S_t) \propto \exp(\lambda V_{f,1}(x^1)) \prod_{h=1}^{H} q(f^h|f^{h+1}, S_t), \tag{3.3}$$

where

$$q(f^h|f^{h+1}, S_t) = \frac{p_0^h(f^h)\exp\left(-\eta L^h(f^h, f^{h+1}; S_t)\right)}{\mathbb{E}_{f^h\sim p_0^h} \exp(-\eta L^h(f^h, f^{h+1}; S_t))}.$$

Algorithm 3 Booster($\mathcal{F}, \eta, \mathcal{D}, \mu_f, T, \lambda$)

1: Draw $g \sim p^{\mu_f}(\cdot|\mathcal{D})$ where the posterior is given by Eqn. (3.6)
2: $\nu_h(x) = \nu_{f,g,h}(x) = \operatorname{argmin}_{\nu\in\Delta_{\mathcal{B}}} \mu_{f,h}^\top g^h(x,\cdot,\cdot)\nu, \forall (x,h)$
3: Return ν.

3.3. The Booster Agent

As aforementioned, the main agent aims to learn an ϵ-close policy. However, given the competing nature of MGs, this task is not feasible if her opponent is naive. Thus, inspired by Jin et al. (2021b); Huang et al. (2021), the second learning agent is set to be the booster agent. As opposed to the main agent, the booster agent does not aim at find her ϵ-close policy. Instead, her goal is to assist the main agent's learning. Specifically, she examines the adopted policy of the main agent and tries to find the best response for it (since the best response is the strongest opponent). In this way, the underlying weakness of the main agent is exploited, which facilitates the learning of the NE. To better illustrate the role of the booster agent, we consider the following decomposition of the regret:

$$\begin{aligned}\mathrm{Reg}(T) = &\underbrace{\left(\sum_{t=1}^{T} V_1^*(x^1) - V_1^{\mu_t,\nu_t}(x^1)\right)}_{\text{main agent}} \\ &+ \underbrace{\left(\sum_{t=1}^{T} V_1^{\mu_t,\nu_t}(x^1) - V_1^{\mu_t,\dagger}(x^1)\right)}_{\text{booster agent}}.\end{aligned} \tag{3.4}$$

The technical advantage of involving $V_1^{\mu_t,\nu_t}(x^1)$ in the main agent part is that we can apply the value-decomposition lemma from Jiang et al. (2017) as in Lemma 1 because (μ_t, ν_t) is the executed policy pair for trajectory collection (see Lemma 1 for details). In this case, the non-negative booster agent part is zero if we can find the best response to μ_t exactly. Motivated by this observation, the booster agent keeps learning to approximate the best response to the given max-player's policy based on the historical trajectories so as to minimize the booster agent part. Due to the different goals, the design philosophy of the booster agent is different from that of the main agent. Especially, she takes a different but also optimistic prior (for the min-player) and a different format of the likelihood.

Optimistic prior of the booster agent. An optimistic prior is adopted for the booster agent, defined as

$$p_0^{\mu}(g) \propto \exp(-\lambda V_{g,1}^{\mu}(x^1)) \prod_{h=1}^{H} p_0^h(g^h). \tag{3.5}$$

Intuitively, the booster agent favors small values for the initial state, which is optimistic for the min-player. The motivation for such an optimistic prior will be clearer after the value decomposition lemma, i.e., Lemma 2, is presented. The reason why we only modify the prior will also be illustrated in Sec. 5.3.

Likelihood for the booster agent. As the booster agent mainly focuses on approximating the best response policy to μ instead of finding NE, a different squared loss function is specified as:

$$L_{\mu}^h(g^h, g^{h+1}; S_t) = \sum_{s=1}^{t} [g^h(x_s^h, a_s^h, b_s^h) - r_s^h - V_{g^{h+1}}^{\mu}(x_s^{h+1})]^2,$$

which can be viewed as a proxy to the squared $\mathcal{T}_h^{\mu}$-Bellman error. Consequently, a corresponding likelihood can be obtained by replacing L^h in Eqn. (3.2) with L_{μ}^h.

Posterior distribution for the booster agent. With the prior and the likelihood, the posterior distribution for the booster agent can be obtained as:

$$p^{\mu}(g|S_t) \propto \exp(-\lambda V_{g,1}^{\mu}(x^1)) \prod_{h=1}^{H} q^{\mu}(g^h|g^{h+1}, S_t), \tag{3.6}$$

where

$$q^{\mu}(g^h|g^{h+1}, S_t) = \frac{p_0^h(g^h) \exp\left(-\eta L_{\mu}^h(g^h, g^{h+1}; S_t)\right)}{\mathbb{E}_{g^h \sim p_0^h} \exp(-\eta L_{\mu}^h(g^h, g^{h+1}; S_t))}.$$

Note that sometimes we also employ the notation $q(g^h|g^{h+1}, \mu, S_t) = q^{\mu}(g^h|g^{h+1}, S_t)$ when we need to use the superscript h.

3.4. The Learning Process

With the main agent and the booster agent specified, the training proceeds as the following. For each episode t, the main agent first samples one $f_t \in \mathcal{F}$ according to the posterior distribution $p(\cdot|S_{t-1})$ and adopts the induced Nash policy as

$$\mu_{t,h}(x) \leftarrow \mu_{f_t,h}(x) := \operatorname{argmax}_{\mu \in \Delta_{\mathcal{A}}} \min_{\nu \in \Delta_{\mathcal{B}}} \mu^{\top} f_t^h(x, \cdot, \cdot)\nu$$

for all $(x, h) \in \mathcal{X} \times [H]$.

Then, the booster agent samples some $g_t \in \mathcal{F}$ from her posterior distribution $p^{\mu_t}(\cdot|S_{t-1})$ computed from S_{t-1} and μ_t. The approximated best response is computed according to g_t and μ_t as

$$\nu_{t,h}(x) \leftarrow \nu_{f_t,g_t,h}(x) = \operatorname{argmin}_{\nu \in \Delta_{\mathcal{B}}} \mu_{f_t,h}^{\top} g_t^h(x, \cdot, \cdot)\nu$$

for all $(x, h) \in \mathcal{X} \times [H]$.

Finally, both players execute (μ_t, ν_t) for episode t, resulting in a trajectory ζ_t. This collected trajectory is then added to S_t and used to compute the policy for the next episode.

4. Sketch of the Main Ideas

In this section, a sketch of the main ideas is provided for the proposed algorithm and the theoretical proof.

4.1. Value-Decomposition Lemmas

It is known that the immediate regret in one episode can be related to the Bellman residuals in the single-agent setting (Jiang et al., 2017), and this technique is well-adopted in the literature (Jin et al., 2021a; Dann et al., 2021; Du et al., 2021). For our setting, with regret decomposed as in Eqn. (3.4), the immediate regrets of the main agent part and the booster agent part can be related to the $\mathcal{T}$-Bellman residuals and the $\mathcal{T}^{\mu_t}$-Bellman residuals, respectively, as we show below.

Lemma 1 (Value decomposition for the main agent.)**.** *Let $\mu = \mu_f$ and ν be an arbitrary policy taken by the min-player. It holds that*

$$V^*(x^1) - V_1^{\mu,\nu}(x^1) \le \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu} \mathcal{E}_h(f^h, f^{h+1}; \zeta) + V^*(x^1) - V_{f,1}(x^1).$$

Lemma 2 (Value decomposition for the booster agent.)**.** *Suppose that $\mu = \mu_f$ is taken by the max-player and g is sampled from the posterior by the booster agent. Let ν be taken as in Sec. 3.4. Then, it holds that*

$$V_1^{\mu,\nu}(x^1) - V_1^{\mu,\dagger}(x^1) = -\sum_{h=1}^{H} \mathbb{E}_{\mu,\nu} \mathcal{E}_h^{\mu}(g^h, g^{h+1}, \zeta) + V_{g,1}^{\mu}(x^1) - V_1^{\mu,\dagger}(x^1).$$

We remark that these two lemmas also account for the extra optimistic terms in the prior distributions. The proofs of these two lemmas are deferred to Appendix E.

4.2. Multi-Agent Decoupling Coefficients

In the previous subsection, we convert the problem of bounding $\mathrm{Reg}(T)$ to bounding the summation of Bellman residuals. However, the posterior distribution is more related to the *squared* Bellman residuals. Therefore, we need some structural information to relate the growth of the cumulative Bellman residuals to the growth of the cumulative squared Bellman residuals. To this end, the multi-agent decoupling coefficient is introduced, which is an extension of the single-agent version in Dann et al. (2021), as follows.

Definition 1 (Multi-agent decoupling coefficient)**.** *Given an $MG(H, \mathcal{X}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r)$, a function class $\mathcal{F}$, a time horizon T, and a parameter $\mu > 0$, the multi-agent decoupling coefficient $dc(\mathcal{F}, MG, T, \mu)$ is defined to be the smallest integer such that*

$$\sum_{h=1}^{H}\sum_{t=1}^{T}\left[\mathbb{E}_{\pi_t}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x^h, a^h, b^h\right)\right]\right] \le \mu\sum_{h=1}^{H}\sum_{t=1}^{T}\left[\sum_{s=1}^{t-1}\left[\mathbb{E}_{\pi_s}\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x^h, a^h, b^h\right)\right]^2\right] + \frac{K}{4\mu},$$

where π_s is a policy pair $(\mu_{f_s}, \nu_{f_s,g_s})$ induced by (f_s, g_s) as introduced in Sec. 3.4. The set of these distributions induced by $f, g \in \mathcal{F}$ is denoted as $\mathcal{D}_{\mathcal{F}}$.

Equipped with the multi-agent decoupling coefficient, it remains to bound the cumulative squared Bellman residuals $\sum_{s=1}^{t-1}\left[\mathbb{E}_{\pi_s}\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x^h, a^h, b^h\right)\right]^2$ by connecting it to the likelihood $L^h_{\mu_{f_t}}(g^h, g^{h+1}; S_{t-1})$ used in the posterior distributions.

4.3. Connection to Likelihood

We focus on the main agent and the booster agent is similar. We consider the $L^h(g_t^h, g_t^{h+1}; \zeta_s)$ (when we only evaluate the loss with only one trajectory, we directly use the notation ζ_s). Taking expectation, we have

$$\mathbb{E}_{\pi_s} L^h(g_t^h, g_t^{h+1}; \zeta_s) = \left[\mathbb{E}_{\pi_s}\mathcal{E}_h(g_t; x^h, a^h, b^h)\right]^2 + \sigma^2, \tag{4.1}$$

where σ^2 is the expectation of $(\mathcal{T}_h f^{h+1}(x_s^h, a_s^h, b_s^h) - r_s^h - V_{f^{h+1}}(x_s^{h+1}))^2$ or the variance, which is hard to deal with. However, the denominator in the likelihood allows us to rewrite the algorithm by replacing $L^h(g_t^h, g_t^{h+1}; \zeta_s)$ with the following excess loss:

$$\Delta L^h(f^h, f^{h+1}; \zeta_s) := L^h(f^h, f^{h+1}; \zeta_s) - (\mathcal{T}_h f^{h+1}(x_s^h, a_s^h, b_s^h) - r_s^h - V_{f^{h+1}}(x_s^{h+1}))^2, \tag{4.2}$$

whose expectation is the desired $\left[\mathbb{E}_{\pi_s}\mathcal{E}_h(g_t; x^h, a^h, b^h)\right]^2$. After resolving the issue of variance, the analysis follows from the online aggregation techniques. However, the completeness assumption is required to analyze the introduced denominator (see the proof of Lemma 11 and Lemma 17).

4.4. More Intuition

We emphasize that the feature of the self-play setting that the learning agent can control both the max-player and the min-player plays a central role in the algorithm design and analysis. This allows us to decompose the regret into two parts as in Eqn. (3.4) and further employ the asymmetric structure to handle two parts. The analysis in the single-agent case essentially relies on the Markov property of transition, (conditional) sub-Gaussianity of the noise of transition, and the fact that the regret in one episode is upper bounded by the sum of Bellman residuals. We note that both the main agent and the booster agent retain these properties separately. Therefore, the techniques from MDP can be applied but with some additional efforts to handle the game nature.

4.5. Complexity of $\mathcal{F}$

For optimization-based algorithms, the complexity of the function class $\mathcal{F}$ is usually characterized through the cardinality $|\mathcal{F}|$ or the covering number (Jiang et al., 2017; Wang et al., 2020; Jin et al., 2021a;b; Huang et al., 2021). On the other hand, the posterior sampling algorithm employs a prior distribution p_0 over $\mathcal{F}$, which allows the algorithm to favor certain parts of it. Accordingly, our theoretical result depends on the complexity of $\mathcal{F}$ through the prior preference, which is characterized by the following quantity.

Definition 2. *For a policy μ_f, $f \in \mathcal{F}$ and for any function $g' \in \mathcal{F}_{h+1}$, we define*

$$\mathcal{F}_h^{\mu_f}(\epsilon, g') = \{g \in \mathcal{F}_h : \sup_{x,a,b} |\mathcal{E}_h^{\mu_f}(g, g'; x, a, b)| \le \epsilon\},$$

containing the functions that have small $\mathcal{T}_h^{\mu_f}$-Bellman error against g' for all state-action pairs. We then define

$$\kappa_\mu(\epsilon) = \sup_{g\in\mathcal{F}} \sum_{h=1}^{H} \ln\left(1/p_0^h(\mathcal{F}_h^{\mu}(\epsilon, g^{h+1}))\right),$$

and

$$\kappa(\epsilon) = \sup_{f\in\mathcal{F}} \kappa_{\mu_f}(\epsilon).$$

Under Assumption 2, it is assumed that $\kappa(\epsilon) < \infty$, which is supported by the following two specific examples.

For the finite function class with completeness, with a uniform prior $p_0^h(f) = 1/|\mathcal{F}_h|$, we have

$$\kappa(\epsilon) \le \sum_{h=1}^{H} \ln|\mathcal{F}_h| = \ln|\mathcal{F}|,$$

due to the realizability assumption. For an infinite function class, by replacing $|\mathcal{F}|$ with its covering number, similar result can also be ontained.

For a d-dimensional parametric models $\mathcal{F}_h = \{g_\theta \in \mathbb{R}^d : \theta \in \Omega_h\}$ where Ω_h is compact, we can generally assume that $\sup_\theta \ln \frac{1}{p_0^h(\{\theta' : \|\theta' - \theta\| \le \epsilon\})} \le d \ln(c'/\epsilon)$ for some constant c' depending on the prior. If we further assume that g_θ is Lipschitz in θ (e.g., linear MG (Xie et al., 2020)), then we can assume that $\ln \frac{1}{p_0^h(\mathcal{F}_h^{\mu_f}(\epsilon, g^{h+1}))} \le c_0 d \ln(c_1/\epsilon)$ for some constants c_0 and c_1 depending on the prior and the Lipschitz constant L. In this case, we have

$$\kappa(\epsilon) \le c_0 H d \ln(c_1/\epsilon).$$

5. Main Results

In this section, we state the main theoretical result of this paper and interpret it using several examples.

5.1. Theoretical Guarantee

We now provide an upper bound for the overall regret.

Theorem 1 (Overall regret). *Let Assumptions 1, 2 and 3 hold. If $\eta\beta^2 \le 0.5$ and $\lambda\beta^2 \ge 1$ hold, and let $dc(\mathcal{F}, MG, T)$ be an upper bound for the $\sup_{\mu \le 1} dc(\mathcal{F}, MG, T, \mu)$, and we further take $\lambda = \sqrt{\frac{T\kappa(\frac{\beta}{T^2})}{\beta^2 dc(\mathcal{F}, MG, T)}}$, $\eta = \frac{1}{4\beta^2}$, then, it holds that*

$$\mathbb{E}\,\mathrm{Reg}(T) \le O\left(\beta\sqrt{dc(\mathcal{F}, MG, T)\kappa(\frac{\beta}{T^2})T} + dc(\mathcal{F}, MG, T)\right).$$

Notably, if the multi-agent decoupling coefficient is provably small, Algorithm 1 admits a $\sqrt{T}$-regret. According to the decomposition in Eqn. (3.4), Theorem 1 can be established once we can bound the main agent part and the booster agent part.

Theorem 2 (Bound of the main agent). *With the same conditions as Theorem 1, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_{S_{t-1}} \mathbb{E}_{f_t \sim p_t} \mathbb{E}_{g_t \sim p_t^{\mu_t}} \left[V_1^*(x^1) - V_1^{\mu_t, \nu_t}(x^1)\right]$$
$$\le O\left(\beta\sqrt{dc(\mathcal{F}, MG, T)\kappa(\frac{\beta}{T^2})T} + dc(\mathcal{F}, MG, T)\right).$$

We then turn to the booster agent and provide an upper bound for the regret induced by approximating the best response policy.

Theorem 3 (Bound of the booster agent). *With the same conditions as Theorem 1, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}_{S_{t-1}} \mathbb{E}_{f_t \sim p_t} \mathbb{E}_{g_t \sim p_t^{\mu_t}} \left[V_1^{\mu_t, \nu_t}(x^1) - V_1^{\mu_t, \dagger}(x^1)\right]$$
$$\le O\left(\beta\sqrt{dc(\mathcal{F}, MG, T)\kappa(\beta/T^2)T} + dc(\mathcal{F}, MG, T)\right).$$

The detailed proofs can be found in the appendix.

5.2. Bounds for the Multi-Agent Decoupling Coefficient

In this subsection, we provide several examples whose multi-agent decoupling coefficient is provably small. The proof can be found in Appendix F.

Linear MG. The first example is the MG with linear function approximation (Xie et al., 2020). In this case, there exists a feature map $\phi(x, a, b) \in \mathbb{R}^d$ and it holds that $r^h(x, a, b) = \phi(x, a, b)^\top \theta_*^h$ and $\mathbb{P}^h(x'|x, a, b) = \phi(x, a, b)^\top \mu_h(x')$ for some unknown $\theta_*^h \in \mathbb{R}^d$ and $\mu_h(\cdot) \in \mathbb{R}^d$ satisfying $\max\{\|\theta_*^h\|, \|\mu_h\|\} \le \sqrt{d}$. We have the following upper bound for the multi-agent decoupling coefficient.

Proposition 1 (Linear MG). *For a d-dimensional MG with $\mathcal{F}_h = \{\phi_h(\cdot,\cdot,\cdot)^\top \theta^h : \|\theta^h\| \le (H + 1 - h)\sqrt{d}\}$ and $\|\phi(x, a, b)\| \le 1, \forall (x, a, b) \in \mathcal{X} \times \mathcal{A} \times \mathcal{B}$, then for all $\mu \le 1$, it holds that*

$$dc(\mathcal{F}, MG, T, \mu) \le 2dH(2 + \ln(2HT)).$$

Note that Jin et al. (2021b) considers a more general setting of linear function approximation whose multi-agent decoupling coefficient is also provably small due to Proposition 3. Also note that as a special case, tabular MG is a linear MG of dimension $d = |\mathcal{X}||\mathcal{A}||\mathcal{B}|$.

Generalized Linear MG. We then consider the generalized linear MG. In this case, we have $(f^h - \mathcal{T}_h^\mu f_{h+1})(x, a, b) = \sigma(\phi(x, a, b)^\top \theta^h)$ for any μ induced by some function in $\mathcal{F}$ and $f \in \mathcal{F}$ where σ is differentiable and strictly increasing. We further assume that $\sigma' \in (c_1, c_2)$ and $\max\{\|\phi(x, a, b)\|, \|\theta^h\|\} \le R$ for some $c_1, c_2, R > 0$.

Proposition 2 (Generalized Linear MG.). *For a generalized linear MG, with $\mathcal{F} = \{(x, a, b) \to \sigma(\phi(x, a, b)^\top \theta : \|\theta\| \le H\sqrt{d}\}$, then for all $\mu \le 1$, it holds that*

$$dc(\mathcal{F}, MG, T, \mu) \le 2dH(c_2^2/c_1^2)(2 + \ln(2HT)).$$

We can also derive an upper bound for the multi-agent decoupling coefficient through multi-agent Bellman Eluder dimension introduced in Jin et al. (2021b).

Proposition 3 (Reduction to multi-agent Bellman Eluder dimension). *Let $\Pi_\mathcal{F} = \mathcal{D}_\mathcal{F}$ be the set of probability measures over $\mathcal{X} \times \mathcal{A} \times \mathcal{B}$ at each step h obtained by following $(\mu_f, \nu_{f,g})$ for some $f, g \in \mathcal{F}$. If for all $\epsilon > 0$ we have*

$$dim_{BE}(\mathcal{F}, \Pi, \epsilon) \le E \ln(1/\epsilon),$$

then for all $\mu \le 1$, the multi-agent decoupling coefficient satisfies:

$$dc(\mathcal{F}, MG, T, \mu) \le 4(1 + \ln(T))EH.$$

Similar to the single-agent case, the multi-agent decoupling coefficient exhibits an additional factor of H due to the formulation of summation over all steps instead of maximum as in the multi-agent Bellman Eluder dimension case. This formulation can offer advantages when the complexity of the function class varies with time steps h. Combining this with Theorem 1, the regret bound of our algorithm matches that of OFU-based algorithms. However, we do remark that the results of Jin et al. (2021b); Huang et al. (2021) are in a high-probability fashion, which is stronger than the bound in expectation.

5.3. Interpretation of Theorem 1

We now illustrate Theorem 1 by concrete examples. The first example is for the finite function classes.

Corollary 4 (Finite function classes with completeness)**.** *Let $\mathcal{F}$ be a finite function class satisfying Assumptions 1, 2 and 3 with $\beta = 2$. Assume that the prior is uniform $p_0^h(f) = 1/|\mathcal{F}_h|$, and $|\mathcal{F}| = \prod_{h=1}^H |\mathcal{F}_h|$. With $\eta = 0.1$ and $\lambda = \sqrt{\frac{T \ln |\mathcal{F}|}{dc(\mathcal{F}, MG, T)}}$, we have*

$$\mathbb{E}\,\mathrm{Reg}(T) = O(\sqrt{dc(\mathcal{F}, MG, T) T \ln(|\mathcal{F}|)}).$$

Note that it is straightforward to generalize this result to the infinite function classes by replacing the cardinality $|\mathcal{F}|$ with its covering number $\mathcal{N}_\infty(\mathcal{F}, \epsilon)$ with an appropriate choice of ϵ. We then illustrate Theorem 1 by considering the MGs with linear function approximation.

Corollary 5 (Linear MG)**.** *For the linear MG, if we assume that the prior is uniform, we have $\kappa(\epsilon) = O(Hd\ln(1/\epsilon))$. With $\eta = \frac{0.4}{H^2}$ and $\lambda = \sqrt{\frac{T\kappa(H/T^2)}{dH^3(1+\ln(2HT))}}$, we have*

$$\mathbb{E}\,\mathrm{Reg}(T) = O(H^2 d\sqrt{T}\ln(HT)).$$

Compared with Xie et al. (2020), our algorithm improves the regret bound for linear MGs by a factor of $\sqrt{d}$. We remark that the improvement is mainly due to the *global* optimism mechanism instead of a step-wise one. Specifically, we add an optimistic term only in the prior distributions at the initial value as in Eqn. (3.1) and Eqn. (3.5). On the contrary, OMVI from Xie et al. (2020) establishes optimism at every step (see lines 8 and 9 of their pseudo code). The main bottleneck is that due to the temporal dependency, OMVI needs to construct uniform concentration for the optimistic bonus function at every step, whose covering number leads to the extra $\sqrt{d}$ factor. See Eqn. 5 and Lemma 18 of Xie et al. (2020) for details.

Recently, Xiong et al. (2022) adopt the dataset splitting trick from MDP (Xie et al., 2021) to resolve this issue in the offline setting where the trajectories are independently collected by some behavior policy. However, their technique cannot apply directly in online setting as the policy used to collect new trajectory depends on the history. Also, we remark that while the OMVI is also computationally efficient, both our posterior sampling algorithm and GOLF of Jin et al. (2021b); Huang et al. (2021) are only information-theoretic. Therefore, it remains open whether we could close this gap by designing computationally efficient algorithm.

6. Conclusion

In this paper, a self-play posterior sampling algorithm is proposed for two-player zero-sum Markov games with general function approximation, which is the first to the best of our knowledge. A new complexity measure, *multi-agent decoupling coefficient*, is introduced to characterize the complexity of function class. Rigorous theoretical analysis showed that the proposed algorithm could achieve comparable regret bounds compared with other OFU-based algorithms for problems with low multi-agent decoupling coefficient, which extends the results in the single-agent RL.

As existing algorithms with general function approximation are computationally inefficient in general, one important direction for future works is to design computationally tractable algorithms for MGs (and MDPs). Another interesting open question is how to extend the posterior sampling algorithms for general-sum Markov games.

Acknowledgements

WX and TZ acknowledge the funding supported by GRF 16201320 and the Hong Kong Ph.D. Fellowship. The CSs acknowledge the funding support by the US National Science Foundation under Grant ECCS- 2029978, ECCS-2033671, and CNS-2002902, and the Bloomberg Data Science Ph.D. Fellowship.

References

Agrawal, P., Chen, J., and Jiang, N. Improved worst-case regret bounds for randomized least-squares value iteration. *arXiv preprint arXiv:2010.12163*, 2020.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dkebiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Brown, N. and Sandholm, T. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Chen, Z., Zhou, D., and Gu, Q. Almost optimal algorithms for two-player Markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021.

Dann, C., Mohri, M., Zhang, T., and Zimmert, J. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Filar, J. and Vrieze, K. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

Jafarnia-Jahromi, M., Jain, R., and Nayyar, A. Learning zero-sum stochastic games with posterior sampling. *arXiv preprint arXiv:2109.03396*, 2021.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713. PMLR, 06–11 Aug 2017.

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, C., Liu, Q., and Yu, T. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021b.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722*, 2016.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. *arXiv preprint arXiv:1406.1853*, 2014.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386. PMLR, 2016.

Perolat, J., Scherrer, B., Piot, B., and Pietquin, O. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*, pp. 1321–1329. PMLR, 2015.

Russo, D. Worst-case regret bounds for exploration via randomized value functions. *arXiv preprint arXiv:1906.02870*, 2019.

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

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

Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Van Handel, R. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Wang, R., Salakhutdinov, R., and Yang, L. F. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *arXiv preprint arXiv:2005.10804*, 2020.

Weisz, G., Amortila, P., and Szepesvári, C. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pp. 1237–1264. PMLR, 2021.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pp. 3674–3682. PMLR, 2020.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

Xiong, Z., Shen, R., and Du, S. S. Randomized exploration is near-optimal for tabular mdp. *arXiv preprint arXiv:2102.09703*, 2021.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964. PMLR, 2020.

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, T. Data dependent concentration bounds for sequential prediction algorithms. In *International Conference on Computational Learning Theory*, pp. 173–187. Springer, 2005.

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *arXiv preprint arXiv:2110.00871*, 2021.

A. Equivalent Algorithms

We will consider a slightly more general posterior sampling algorithm with an extra parameter $\alpha \in (0, 1]$. We recall that the posterior defined in Eqn. (3.3) is

$$p(f|S_t) \propto \exp(\lambda V_{f,1}(x^1)) \prod_{h=1}^{H} q(f^h|f^{h+1}, S_t),$$

where

$$q(f^h|f^{h+1}, S_t) = \frac{p_0^h(f^h) \exp\left(-\eta L^h(f^h, f^{h+1}; S_t)\right)}{\mathbb{E}_{f^h \sim p_0^h} \exp(-\eta L^h(f^h, f^{h+1}; S_t))}.$$

Equivalently, we may consider the excess loss

$$\begin{aligned}\Delta L^h(f^h, f^{h+1}; \zeta_s) =& (f^h(x_s^h, a_s^h, b_s^h) - r_s^h - V_{f^{h+1}}(x_s^{h+1}))^2 \\ &- (\mathcal{T}_h f^{h+1}(x_s^h, a_s^h, b_s^h) - r_s^h - V_{f^{h+1}}(x_s^{h+1}))^2,\end{aligned} \tag{A.1}$$

where we employ the notation that $\zeta_s = \{[x_s^h, a_s^h, b_s^h, r_s^h]\}_{h=1}^H$. We then define the potential function as

$$\begin{aligned}\Phi_t^h(f) =& -\ln p_0^h\left(f^h\right) + \alpha\eta \sum_{s=1}^{t-1} \Delta L^h\left(f^h, f^{h+1}; \zeta_s\right) \\ &+ \alpha \ln \mathbb{E}_{\tilde{f}^h \sim p_0^h} \exp\left(-\eta \sum_{s=1}^{t-1} \Delta L^h\left(\tilde{f}^h, f^{h+1}; \zeta_s\right)\right),\end{aligned} \tag{A.2}$$

where $\alpha \in (0, 1]$ is the extra parameter to facilitate the proof. We also define

$$\Delta f^1(x^1) = V_{f,1}(x^1) - V_1^*(x^1).$$

Then, we obtain a generalized posterior distribution on $\mathcal{F}$:

$$\hat{p}_t(f) \propto \exp\left(-\sum_{h=1}^{H} \Phi_t^h(f) + \lambda \Delta f^1(x^1)\right), \tag{A.3}$$

where it is equivalent to the posterior given in Eqn. (3.3) when $\alpha = 1$.

We then recall the posterior distribution of the booster agent defined in Eqn. (3.6) is given by

$$p^\mu(g|S_t) \propto \exp(-\lambda V_{g,1}^\mu(x^1)) \prod_{h=1}^{H} q^\mu(g^h|g^{h+1}, S_t),$$

where

$$q^\mu(g^h|g^{h+1}, S_t) = \frac{p_0^h(g^h) \exp\left(-\eta L_\mu^h(g^h, g^{h+1}; S_t)\right)}{\mathbb{E}_{g^h \sim p_0^h} \exp(-\eta L_\mu^h(g^h, g^{h+1}; S_t))}.$$

Similarly, we define the excess loss for the booster agent:

$$\begin{aligned}\Delta L_\mu^h(g^h, g^{h+1}; \zeta_s) =& (g^h(x_s^h, a_s^h, b_s^h) - r_s^h - V_{g^{h+1}}^\mu(x_s^{h+1}))^2 \\ &- (\mathcal{T}_h^\mu g^{h+1}(x_s^h, a_s^h, b_s^h) - r_s^h - V_{g^{h+1}}^\mu(x_s^{h+1}))^2.\end{aligned} \tag{A.4}$$

and

$$\Delta g_\mu^1(x^1) = V_1^{\mu,\dagger}(v^1) - V_{g,1}^\mu(x^1),$$

and use the following notation (with slight abuse of notation) for the potential function:

$$\begin{aligned}\Phi_t^h(g, \mu) =& -\ln p_0^h\left(g^h\right) + \alpha\eta \sum_{s=1}^{t-1} \Delta L_\mu^h\left(g^h, g^{h+1}; \zeta_s\right) \\ &+ \alpha \ln \mathbb{E}_{\tilde{g}^h \sim p_0^h} \exp\left(-\eta \sum_{s=1}^{t-1} \Delta L_\mu^h\left(\tilde{g}^h, g^{h+1}; \zeta_s\right)\right),\end{aligned} \tag{A.5}$$

since the analyses for Algorithm 2 and Algorithm 3 are separate so the meaning of $\Phi_t^h(\cdot)$ will be clear from the context. Finally, we obtain a generalized posterior function for the booster agent:

$$\hat{p}_t^{\mu}(g) \propto \exp\left(-\sum_{h=1}^{H} \Phi_t^h(g,\mu) + \lambda \Delta g_{\mu}^1(x^1)\right). \tag{A.6}$$

The main motivation to use $\Delta L^h(\cdot)$ ($\Delta L_{\mu}^h(\cdot)$) is that the variance will be cancelled during our theoretical analysis as it is equivalent to the case where we know the Bellman operator. This is possible because the novel denominator term is introduced in the likelihood as in Dann et al. (2021).

B. Useful Lemmas and Additional Notations

In this section, we provide several useful lemmas and additional notations that are useful later. We start with the following definitions. First, we further define a quantity similar to Definition 2, which will be used for the analysis of the main agent.

Definition 3. *For any $f' \in \mathcal{F}_{h+1}$, we define the set*

$$\mathcal{F}_h(\epsilon, f') := \{f \in \mathcal{F}_h : \sup_{x,a,b} |\mathcal{E}_h(f, f'; x, a, b)| \leq \epsilon\}$$

containing the functions that have small $\mathcal{T}_h$-Bellman error against f' for all state-action pairs. We then define the quantity

$$\kappa_1(\epsilon) = \sup_{f \in \mathcal{F}} \sum_{h=1}^{H} \ln \frac{1}{p_0^h(\mathcal{F}_h(\epsilon, f^{h+1}))},$$

which is the probability assigned by the prior to functions that approximately satisfy the Bellman equation w.r.t. f for all state-action pair.

Note that $\kappa_1(\epsilon) \leq \kappa(\epsilon)$ because

$$\kappa_1(\epsilon) = \sup_{g \in \mathcal{F}} \sum_{h=1}^{H} \ln \frac{1}{p_0^h(\mathcal{F}_h^{\mu_g}(\epsilon, g^{h+1}))} \leq \sup_{f \in \mathcal{F}} \sup_{g \in \mathcal{F}} \sum_{h=1}^{H} \ln \frac{1}{p_0^h(\mathcal{F}_h^{\mu_f}(\epsilon, g^{h+1}))} = \kappa(\epsilon).$$

Definition 4. *For $\alpha \in (0,1)$, we also use the notations:*

$$\kappa_1^h(\alpha, \epsilon) = (1-\alpha) \ln \mathbb{E}_{f^{h+1} \sim p_0^{h+1}} p_0^h(\mathcal{F}_h(\epsilon, f^{h+1}))^{-\alpha/(1-\alpha)},$$

and $\kappa_1^h(1,\epsilon) = \lim_{\alpha \to 1^-} \kappa^h(\alpha, \epsilon)$ where it holds that

$$\kappa_1^h(1,\epsilon) = \sup_{f^{h+1} \in \mathcal{F}_{h+1}} \ln \frac{1}{p_0^h\left(\mathcal{F}_h\left(\epsilon, f^{h+1}\right)\right)} < \infty,$$

and

$$\kappa_1(\epsilon) = \sum_{h=1}^{H} \kappa_1^h(1,\epsilon) \leq \kappa(\epsilon).$$

Similarly, we define

$$\kappa_{\mu}^h(\alpha, \epsilon) = (1-\alpha) \ln \mathbb{E}_{f^{h+1} \sim p_0^{h+1}} p_0^h(\mathcal{F}_h^{\mu}(\epsilon, f^{h+1}))^{-\alpha/(1-\alpha)},$$

and $\kappa_{\mu}^h(1,\epsilon) = \lim_{\alpha \to 1^-} \kappa_{\mu}^h(\alpha, \epsilon)$, Then, it holds that

$$\kappa_{\mu}^h(1,\epsilon) = \sup_{f^{h+1} \in \mathcal{F}_{h+1}} \ln \frac{1}{p_0^h(\mathcal{F}_h^{\mu}(\epsilon, f^{h+1}))} < \infty,$$

and

$$\kappa_{\mu}(\epsilon) = \sum_{h=1}^{H} \kappa_{\mu}^h(1,\epsilon) \leq \kappa(\epsilon).$$

Lemma 3. *For any fixed $g \in \mathcal{F}$ and max-player's policy $\mu := \mu_f$ for some $f \in \mathcal{F}$, we define a random variable for all s and h as follows:*

$$\xi_s^h\left(g^h, g^{h+1}, \zeta_s\right) = -2\eta\Delta L_\mu^h\left(g^h, g^{h+1}, \zeta_s\right) - \ln \mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h, a_s^h)} \exp\left(-2\eta\Delta L_\mu^h\left(g^h, g^{h+1}, \zeta_s\right)\right).$$

Then, for all h, we have

$$\mathbb{E}_{S_{t-1}} \exp\left(\sum_{s=1}^{t-1} \xi_s^h\left(g^h, g^{h+1}, \zeta_s\right)\right) = 1.$$

A special case is that $f = g$ where we have

$$\Delta L_{\mu_f}^h(f^h, f^{h+1}, \zeta_s) = \Delta L^h(f^h, f^{h+1}, \zeta_s).$$

Proof. This lemma is from Zhang (2005) and is also proved in Dann et al. (2021). □

Lemma 4. *Let ν be a probability distribution. Then, $\mathbb{E}_\nu f - H(\nu)$ is minimized at $\nu(x) \propto \exp(-f(x))$.*

Proof. This is a corollary of Gibbs variational principle whose proof can be found in Van Handel (2014), Lemma 4.10. □

Using Lemma 4, we can obtain the following key lemma as used in Dann et al. (2021).

Lemma 5. *It holds that*

$$\begin{aligned}
\mathbb{E}_{f\sim\hat{p}_t}\left(\sum_{h=1}^H \Phi_t^h(f) - \lambda\Delta f^1\left(x^1\right) + \ln\hat{p}_t(f)\right) &= \inf_p \mathbb{E}_{f\sim p(\cdot)}\left(\sum_{h=1}^H \Phi_t^h(f) - \lambda\Delta f^1\left(x^1\right) + \ln p(f)\right); \\
\mathbb{E}_{g\sim\hat{p}_t^\mu}\left(\sum_{h=1}^H \Phi_t^h(g,\mu) - \lambda\Delta g_\mu^1\left(x^1\right) + \ln\hat{p}_t^\mu(g)\right) &= \inf_p \mathbb{E}_{g\sim p(\cdot)}\left(\sum_{h=1}^H \Phi_t^h(g,\mu) - \lambda\Delta g_\mu^1\left(x^1\right) + \ln p(g)\right),
\end{aligned} \tag{B.1}$$

where we remark that the definitions of $\Phi_t^h(\cdot)$ in two equations are different.

In what follows, we derive a lower bound of LHS of Eqn. (B.1), and an upper bound of RHS of Eqn. (B.1) for the proof of Theorems 2 and 3.

C. Proof of the Theorem 2

In this section, we provide the proof for Theorem 2. The proof provided in this section basically follows the same line of that of single-agent RL because essentially the algorithms employ the same properties of the problem as discussed in Section 4 and for the main agent, and the Bellman residuals is free of the min-player's policy.

Lemma 6. *For all functions $f \in \mathcal{F}$, we have*

$$\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h, a_s^h, b_s^h)} \Delta L^h\left(f^h, f^{h+1}, \zeta_s\right) = \left(\mathcal{E}_h\left(f; x_s^h, a_s^h, b_s^h\right)\right)^2$$

and

$$\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h, a_s^h, b_s^h)} \Delta L^h\left(f^h, f^{h+1}, \zeta_s\right)^2 \le \frac{4\beta^2}{3}\left(\mathcal{E}_h\left(f; x_s^h, a_s^h, b_s^h\right)\right)^2$$

Proof. We define the random variable

$$Z = f^h(x_s^h, a_s^h, b_s^h) - r_s^h - V_{f,h+1}(x_s^{h+1}).$$

Let $\mathbb{E}$ be conditioned on $[x_s^h, a_s^h, b_s^h]$. Then, the randomness is from the state transition and we have

$$\mathbb{E}Z = \mathcal{E}_h(f; x_s^h, a_s^h, b_s^h).$$

We also have

$$\Delta L^h(f^h, f^{h+1}, \zeta_s) = Z^2 - (Z - \mathbb{E}Z)^2.$$

and

$$\mathbb{E}[Z^2 - (Z - \mathbb{E}Z)^2] = \mathbb{E}Z^2 - \text{var}(Z) = (\mathbb{E}Z)^2 = (\mathcal{E}_h(f; x_s^h, a_s^h, b_s^h))^2.$$

Also note that $Z \in [-\beta, \beta - 1]$ and $\max Z - \min Z \leq \beta$ if it is conditioned on $[x_s^h, a_s^h, b_s^h]$, this implies that

$$\mathbb{E}(Z^2 - (Z - \mathbb{E}Z)^2)^2 \leq \frac{4}{3}\beta^2(\mathbb{E}Z)^2.$$

□

Lemma 7. *If the learning rate η is sufficiently small such that $\eta\beta^2 \leq 0.8$, then for all functions $f \in \mathcal{F}$, we have*

$$\begin{aligned}
&\ln \mathbb{E}_{x_s^{h+1} \sim \mathbb{P}^h(\cdot|x_s^h, a_s^h, b_s^h)} \exp\left(-\eta \Delta L^h\left(f^h, f^{h+1}, \zeta_s\right)\right) \\
&\quad \leq \mathbb{E}_{x_s^{h+1} \sim \mathbb{P}^h(\cdot|x_s^h, a_s^h, b_s^h)} \exp\left(-\eta \Delta L^h\left(f^h, f^{h+1}, \zeta_s\right)\right) - 1 \\
&\quad \leq -0.25\eta \left(\mathcal{E}_h\left(f; x_s^h, a_s^h, b_s^h\right)\right)^2.
\end{aligned}$$

Proof. With $\eta\beta^2 \leq 0.8$, for all $f \in \mathcal{F}$, we have

$$-\eta \Delta L^h(f^h, f^{h+1}, \zeta_s) \leq 0.8.$$

This implies that

$$\begin{aligned}
&\exp\left(-\eta \Delta L^h(f^h, f^{h+1}, \zeta_s)\right) \\
&\leq 1 - \eta \Delta L^h(f^h, f^{h+1}, \zeta_s) + 0.67\eta^2 \Delta L^h(f^h, f^{h+1}, \zeta_s)^2,
\end{aligned}$$

where we use the fact that $\psi(z) = (e^z - 1 - z)/z^2$ is increasing in z and $\psi(0.8) < 0.67$. Therefore, we have

$$\begin{aligned}
&\ln \mathbb{E}_{x_s^{h+1} \sim \mathbb{P}^h(\cdot|x_s^h, a_s^h, b_s^h)} \exp\left(-\eta \Delta L^h\left(f^h, f^{h+1}, \zeta_s\right)\right) \\
&\leq \mathbb{E}_{x_s^{h+1} \sim \mathbb{P}^h(\cdot|x_s^h, a_s^h, b_s^h)} \exp\left(-\eta \Delta L^h\left(f^h, f^{h+1}, \zeta_s\right)\right) - 1 \\
&\leq \mathbb{E}_{x_s^{h+1} \sim \mathbb{P}^h(\cdot|x_s^h, a_s^h, b_s^h)} - \eta \Delta L^h(f^h, f^{h+1}, \zeta_s) + 0.67\eta^2 \Delta L^h(f^h, f^{h+1}, \zeta_s)^2 \\
&\leq -0.25\eta(\mathcal{E}_h(f^h, f^{h+1}, \zeta_s)^2,
\end{aligned}$$

where the first inequality is because $\ln z \leq z - 1$ and the last inequality is because Lemma 6 and $(\frac{4}{3}\eta b^2 0.67) \leq 0.75$. □

Lemma 8. *It holds that*

$$\inf_p \mathbb{E}_{S_{t-1}} \mathbb{E}_{f \sim p(\cdot)} \left[\sum_{h=1}^H \Phi_t^h(f) - \lambda \Delta f^1\left(x^1\right) + \ln p(f)\right] \leq \lambda\epsilon + 4\alpha\eta(t-1)H\epsilon^2 - \sum_{h=1}^H \ln p_0^h\left(\mathcal{F}_h\left(\epsilon, Q_{h+1}^*\right)\right).$$

Proof. Consider any fixed $f \in \mathcal{F}$. For any $\tilde{f}^h \in \mathcal{F}^h$ that depends on S_{s-1}, we obtain from Lemma 7 that

$$\mathbb{E}_{\zeta_s} \exp\left(-\eta \Delta L^h\left(\tilde{f}^h, f^{h+1}, \zeta_s\right)\right) - 1 \leq -0.25\eta \mathbb{E}_{\zeta_s}\left(\tilde{f}^h(x, a) - \mathcal{T}_h f^{h+1}(x, a, b)\right)^2 \leq 0.$$

We let

$$W_t^h := \mathbb{E}_{S_t} \mathbb{E}_{f \sim p(\cdot)} \ln \mathbb{E}_{\tilde{f}^h \sim p_0^h} \exp\left(-\eta \sum_{s=1}^t \Delta L^h(\tilde{f}^h, f^{h+1}, \zeta_s)\right),$$

and recall that

$$\hat{q}_t^h\left(\tilde{f}^h \mid f^{h+1}, S_{t-1}\right) = \frac{p_0^h(\tilde{f}^h) \exp\left(-\eta \sum_{s=1}^{t-1} \Delta L^h\left(\tilde{f}^h, f^{h+1}, \zeta_s\right)\right)}{\mathbb{E}_{\tilde{f}' \sim p_0^h} \exp\left(-\eta \sum_{s=1}^{t-1} \Delta L^h\left(\tilde{f}', f^{h+1}, \zeta_s\right)\right)}.$$

We have

$$W_s^h - W_{s-1}^h = \mathbb{E}_{S_s} \mathbb{E}_{f \sim p(\cdot)} \ln \mathbb{E}_{\tilde{f}^h \sim p_0^h} \frac{\exp\left(-\eta \sum_{t=1}^{s-1} \Delta L^h\left(\tilde{f}^h, f^{h+1}, \zeta_t\right)\right)}{\mathbb{E}_{\tilde{f}' \sim p_0^h} \exp\left(-\eta \sum_{t=1}^{s-1} \Delta L^h\left(\tilde{f}', f^{h+1}, \zeta_t\right)\right)} \exp\left(-\eta \Delta L^h\left(\tilde{f}^h, f^{h+1}, \zeta_s\right)\right)$$

$$= \mathbb{E}_{S_s}\mathbb{E}_{f\sim p(\cdot)} \ln \mathbb{E}_{\tilde{f}^h\sim \hat{q}_s^h(\cdot|f^{h+1},S_{s-1})} \exp\left(-\eta \Delta L^h\left(\tilde{f}^h, f^{h+1}, \zeta_s\right)\right)$$
$$\leq \mathbb{E}_{S_s}\mathbb{E}_{f\sim p(\cdot)} \left(\mathbb{E}_{\tilde{f}^h\sim \hat{q}_s^h(\cdot|f^{h+1},S_{s-1})} \exp\left(-\eta \Delta L^h\left(\tilde{f}^h, f^{h+1}, \zeta_s\right)\right) - 1\right) \leq 0$$

where we use $\ln z \leq z-1$. By $W_0^h = 0$, we know that

$$W_t^h = W_0^h + \sum_{s=1}^{t}[W_s^h - W_{s-1}^h] \leq 0,$$

equivalently,

$$\mathbb{E}_{S_t}\mathbb{E}_{f\sim p(\cdot)} \ln \mathbb{E}_{\tilde{f}^h\sim p_0^h} \exp\left(-\eta \sum_{s=1}^{t} \Delta L^h(\tilde{f}^h, f^{h+1}, \zeta_s)\right) \leq 0.$$

This implies that for any $p(\cdot)$, we have

$$\begin{aligned}
&\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim p(\cdot)}\left[\sum_{h=1}^{H}\Phi_t^h(f) - \lambda\Delta f^1\left(x^1\right) + \ln p(f)\right] \\
&= \mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim p(\cdot)}\left[-\lambda\Delta f^1\left(x^1\right) + \alpha\eta\sum_{h=1}^{H}\sum_{s=1}^{t-1}\Delta L^h\left(f^h, f^{h+1}, \zeta_s\right)\right. \\
&\qquad \left. +\alpha\sum_{h=1}^{H}\ln \mathbb{E}_{\tilde{f}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L^h\left(\tilde{f}^h, f^{h+1}, \zeta_s\right)\right) + \ln\frac{p(f)}{p_0(f)}\right] \\
&\leq \mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim p(\cdot)}\left[-\lambda\Delta f^1\left(x^1\right) + \sum_{h=1}^{H}\alpha\eta\sum_{s=1}^{t-1}\left(\mathcal{E}_h\left(f; x_s^h, a_s^h, b_s^h\right)\right)^2 + \ln\frac{p(f)}{p_0(f)}\right],
\end{aligned}$$

where we use the definition of the potential function in first equality. Since $p(\cdot)$ is arbitrary, we can take $f^h \in \mathcal{F}_h(\epsilon, Q_{h+1}^*)$ for all $h \in [H]$. We then have

$$|f^h(x,a,b) - Q_h^*(x,a,b)| = |f^h(x,a,b) - \mathcal{T}Q_{h+1}^*(x,a,b)| \leq \epsilon,$$

for all $(x,a,b,h) \in \mathcal{X}\times\mathcal{A}\times\mathcal{B}\times[H]$. Then, we have

$$|\mathcal{E}_h(f;x,a,b)| \leq |f^h(x,a,b) - Q_h^*(x,a,b)| + \sup_{x'}|V_{f,h+1}(x') - V_{h+1}^*(x')| \leq 2\epsilon,$$

where we use

$$\begin{aligned}
|V_{f,h+1}(x') - V_{h+1}^*(x')| &= |\sup_{\mu}\inf_{\nu}\mathbb{D}_{\mu,\nu}f^{h+1}(x') - \sup_{\mu}\inf_{\nu}\mathbb{D}_{\mu,\nu}Q_{h+1}^*(x')| \\
&\leq \sup_{\mu}\sup_{\nu}|\mathbb{D}_{\mu,\nu}(f^{h+1}(x') - Q_{h+1}^*(x'))| \leq \epsilon,
\end{aligned}$$

where the first inequality is because of

$$|\inf_A f - \inf_A g| \leq \sup_A |f-g|.$$

By taking $p(f) = p_0(f)I(f\in\mathcal{F}(\epsilon))/p_0(\mathcal{F}(\epsilon))$, with $\mathcal{F}(\epsilon) = \prod_h \mathcal{F}_h(\epsilon, Q_{h+1}^*)$, we obtain the desired result. □

Lemma 9. *It holds that*

$$\begin{aligned}
\mathbb{E}_{f\sim\hat{p}_t(f)}\ln\hat{p}_t(f) \geq& \alpha\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t(f) + (1-\alpha)\mathbb{E}_{f\sim\hat{p}_t}\sum_{h=1}^{H}\ln\hat{p}_t\left(f^h\right) \\
\geq& \frac{\alpha}{2}\sum_{h=1}^{H}\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^h, f^{h+1}\right) \\
&+ (1-0.5\alpha)\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^1\right) + (1-\alpha)\sum_{h=2}^{H}\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^h\right).
\end{aligned} \tag{C.1}$$

Proof. To show the first inequality, we just subtract all terms of RHS from LHS to see that it is a KL-divergence which is non-negative. The second inequality is equivalent to

$$\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t(f)\geq 0.5\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^1\right)+0.5\sum_{h=1}^{H}\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^h,f^{h+1}\right).$$

This follows from

$$0.5\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t(f)\geq 0.5\sum_{h=1}^{H}\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^h,f^{h+1}\right)I(h\text{ is a odd number })$$

and

$$0.5\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t(f)\geq 0.5\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^1\right)+0.5\sum_{h=1}^{H}\mathbb{E}_{f\sim\hat{p}_t}\ln\hat{p}_t\left(f^h,f^{h+1}\right)I(h\text{ is an even number })$$

which is a result of the non-negativity of mutual information. □

Lemma 10. *It holds that*

$$\begin{aligned}
&\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\left(\sum_{h=1}^{H}\Phi_t^h(f)-\lambda\Delta f^1\left(x^1\right)+\ln\hat{p}_t(f)\right)\\
&\quad\geq\underbrace{\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\left[-\lambda\Delta f^1\left(x^1\right)+(1-0.5\alpha)\ln\frac{\hat{p}_t\left(f^1\right)}{p_0^1\left(f^1\right)}\right]}_{A}\\
&\qquad+\sum_{h=1}^{H}\underbrace{0.5\alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\left[\eta\sum_{s=1}^{t-1}2\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)+\ln\frac{\hat{p}_t\left(f^h,f^{h+1}\right)}{p_0^h\left(f^h\right)p_0^{h+1}\left(f^{h+1}\right)}\right]}_{B_h}\\
&\qquad+\sum_{h=1}^{H}\underbrace{\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\left[\alpha\ln\mathbb{E}_{\tilde{f}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L^h\left(\tilde{f}^h,f^{h+1},\zeta_s\right)\right)+(1-\alpha)\ln\frac{\hat{p}_t\left(f^{h+1}\right)}{p_0^{h+1}\left(f^{h+1}\right)}\right]}_{C_h}.
\end{aligned}\tag{C.2}$$

Proof. We use the definition of the potential function and apply Lemma 9. The desired result follows from some calculations. □

Lemma 11. *If $\eta\beta^2\leq 0.4$, it holds that*

$$A\geq-\lambda\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\Delta f_t^1(x^1),\tag{C.3}$$

$$B_h\geq 0.25\alpha\eta\sum_{s=1}^{t-1}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\mathbb{E}_{\pi_s}\left(\mathcal{E}_h\left(f;x_s^h,a_s^h,b_s^h\right)\right)^2\tag{C.4}$$

$$C_h\geq-\alpha\eta\epsilon(2b+\epsilon)(t-1)-\kappa_1^h(\alpha,\epsilon).\tag{C.5}$$

Proof. The bound of A comes from the non-negativity of KL-divergence and $\alpha\in(0,1]$. To prove the lower bound of B_h, we define

$$\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)=-2\eta\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)-\ln\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)}\exp\left(-2\eta\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)\right).$$

Then, for all $h\in[H]$, we have

$$\mathbb{E}_{S_{t-1}}\exp\left(\sum_{s=1}^{t-1}\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)\right)=1,$$

according to Lemma 3. Then, by Lemma 4, we have

$$\begin{aligned}
&\mathbb{E}_{f\sim\hat{p}_t}\left[\sum_{s=1}^{t-1}-\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)+\ln\frac{\hat{p}_t\left(f^h,f^{h+1}\right)}{p_0^h\left(f^h\right)p_0^{h+1}\left(f^{h+1}\right)}\right]\\
&\geq\inf_p\mathbb{E}_{f\sim p}\left[\sum_{s=1}^{t-1}-\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)+\ln\frac{p\left(f^h,f^{h+1}\right)}{p_0^h\left(f^h\right)p_0^{h+1}\left(f^{h+1}\right)}\right]\\
&=-\ln\mathbb{E}_{f^h\sim p_0^h}\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}\exp\left(\sum_{s=1}^{t-1}\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)\right),
\end{aligned}$$

where we use the fact that Lemma 4 implies that the $\inf$ is achieved at

$$p\left(f^h,f^{h+1}\right)\propto p_0^h\left(f^h\right)p_0^{h+1}\left(f^{h+1}\right)\exp\left(\sum_{s=1}^{t-1}\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)\right),$$

and the expectation is equal to

$$-\mathbb{E}_{p(f^h,f^{h+1})}\sum_{s=1}^{t-1}\xi_s^h(f^h,f^{h+1},\zeta_s)+\mathbb{E}_{p(f^h,f^{h+1})}\ln\frac{\exp(\sum_{s=1}^{t-1}\xi_s^h(f^h,f^{h+1},\zeta_s))}{c}=-\ln c$$

where $c=\mathbb{E}_{f^h\sim p_0^h}\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}\exp\left(\sum_{s=1}^{t-1}\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)\right)$ is the normalized constant. It then follows that

$$\begin{aligned}
&\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\left[\sum_{s=1}^{t-1}-\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)+\ln\frac{\hat{p}_t\left(f^h,f^{h+1}\right)}{p_0^h\left(f^h\right)p_0^{h+1}\left(f^{h+1}\right)}\right]\\
&\geq-\mathbb{E}_{S_{t-1}}\ln\mathbb{E}_{f^h\sim p_0^h}\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}\exp\left(\sum_{s=1}^{t-1}\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)\right)\\
&\geq-\ln\mathbb{E}_{f^h\sim p_0^h}\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}\mathbb{E}_{S_{t-1}}\exp\left(\sum_{s=1}^{t-1}\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)\right)=0,
\end{aligned}$$

where we use the above result in the first inequality and use the convexity of $-\ln(\cdot)$ in the last inequality. We then have

$$\begin{aligned}
B_h&=0.5\alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\left[\eta\sum_{s=1}^{t-1}2\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)+\ln\frac{\hat{p}_t\left(f^h,f^{h+1}\right)}{p_0^h\left(f^h\right)p_0^{h+1}\left(f^{h+1}\right)}\right]\\
&\geq 0.5\alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\sum_{s=1}^{t-1}-\ln\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)}\exp\left(-2\eta\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)\right)\\
&\geq-0.5\alpha\eta\sum_{s=1}^{t-1}\frac{1}{2}(\mathcal{E}_h(f;x_s^h,a_s^h,r_s^h))^2,
\end{aligned}$$

where we use the definition of $\xi_s^h\left(f^h,f^{h+1},\zeta_s\right)$ in the first inequality and we use Lemma 7 in the last step.

We now turn to the lower bound of C_h. We have

$$\begin{aligned}
&\mathbb{E}_{f\sim\hat{p}_t}\left[\alpha\ln\mathbb{E}_{\tilde{f}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L^h\left(\tilde{f}^h,f^{h+1},\zeta_s\right)\right)+(1-\alpha)\ln\frac{\hat{p}_t\left(f^{h+1}\right)}{p_0^{h+1}\left(f^{h+1}\right)}\right]\\
&\geq(1-\alpha)\inf_{p^h}\mathbb{E}_{f\sim p^h}\left[\frac{\alpha}{1-\alpha}\ln\mathbb{E}_{\tilde{f}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L^h\left(\tilde{f}^h,f^{h+1},\zeta_s\right)\right)+\ln\frac{p^h\left(f^{h+1}\right)}{p_0^{h+1}\left(f^{h+1}\right)}\right]\\
&=-(1-\alpha)\ln\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}\left(\mathbb{E}_{f^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)\right)\right)^{-\alpha/(1-\alpha)},
\end{aligned}$$

where we use the fact that the $\inf$ is achieved at

$$p^h\left(f^{h+1}\right) \propto p_0^{h+1}\left(f^{h+1}\right)\left(\mathbb{E}_{f^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)\right)\right)^{-\alpha/(1-\alpha)}.$$

We now consider a fixed $f^h\in\mathcal{F}_h(\epsilon,f^{h+1})$. It holds that

$$\left|\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)\right| \le \left(\mathcal{E}_h\left(f,x_s^h,a_s^h\right)\right)^2 + 2\beta\left|\mathcal{E}_h\left(f,x_s^h,a_s^h\right)\right| \le \epsilon(2\beta+\epsilon)$$

To show this, we recall the definition

$$\begin{aligned}\Delta L^h\left(f^h,f^{h+1};\zeta_s\right) = &\left(f^h\left(x_s^h,a_s^h,b_s^h\right)-r_s^h-V_{f,h+1}\left(x_s^{h+1}\right)\right)^2\\ &-\left(\mathcal{T}_h f^{h+1}\left(x_s^h,a_s^h,b_s^h\right)-r_s^h-V_{f,h+1}\left(x_s^{h+1}\right)\right)^2,\end{aligned}$$

and we subtract and add $\mathcal{T}_h f^{h+1}(x_s^h,a_s^h)$ inside the first term to obtain

$$\Delta L^h(f^h,f^{h+1},\zeta_s) = \mathcal{E}_h(f,x_s^h,a_s^h)^2 + 2\mathcal{E}_h(f,x_s^h,a_s^h)(\mathcal{T}_h^* f^{h+1}(x_s^h,a_s^h)-r_s^h-f^{h+1}(x_s^{h+1})).$$

It follows that

$$\mathbb{E}_{f^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L^h\left(f^h,f^{h+1},\zeta_s\right)\right) \le p_0^h\left(\mathcal{F}_h\left(\epsilon,f^{h+1}\right)\right)\exp(-\eta(t-1)(2\beta+\epsilon)\epsilon).$$

Thus, we have

$$\begin{aligned}C_h &\ge \alpha\mathbb{E}_{S_{t-1}}\ln\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}p_0^h(\mathcal{F}_h(\epsilon,f^{h+1}))\exp(-\eta(t-1)(2\beta+\epsilon)\epsilon)\\ &= -\alpha\eta\epsilon(2\beta+\epsilon)(t-1)+\alpha\mathbb{E}_{S_{t-1}}\ln\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}p_0^h(\mathcal{F}_h(\epsilon,f^{h+1}))\\ &\ge -\alpha\eta\epsilon(2\beta+\epsilon)(t-1)-\kappa_1^h(\alpha,\epsilon)\end{aligned}$$

where we use the definition

$$\kappa_1^h(\alpha,\epsilon) = (1-\alpha)\ln\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}p_0^h\left(\mathcal{F}_h\left(\epsilon,f^{h+1}\right)\right)^{-\alpha/(1-\alpha)}.$$

□

We are ready to prove Theorem 2.

Proof of Theorem 2. Let π_t denote the distribution induced by $\mu_t\times\nu_t$ and define

$$\delta_t^h = \lambda\mathcal{E}_h(f_t;x_t^h,a_t^h,b_t^h) - 0.25\alpha\eta\sum_{s=1}^{t-1}\mathbb{E}_{\pi_s}\left(\mathcal{E}_h(f_t;x_t^h,a_t^h,b_t^h)\right)^2.$$

Then, we have

$$\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\mathbb{E}_{\zeta_t\sim\pi_t}\sum_{h=1}^{H}\delta_t^h \le \frac{\lambda^2}{\alpha\eta}dc(\mathcal{F},MG,T,0.25\alpha\eta/\lambda).$$

For arbitrary ν_t induced by μ_{f_t} and g_t, according to the value-decomposition Lemma 1 we have

$$\begin{aligned}
&\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\lambda(V_1^*(x^1)-V_1^{\mu_t,\nu_t}(x^1))-\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\mathbb{E}_{\zeta_t\sim\pi_t}\sum_{h=1}^{H}\delta_t^h\\
&\leq -\lambda\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\Delta f_t^1(x^1)+0.25\alpha\eta\sum_{h=1}^{H}\sum_{s=1}^{t-1}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{\pi_s}\left(\mathcal{E}_h(f_t;x_t^h,a_t^h,b_t^h)\right)^2\\
&\leq \mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\left(\sum_{h=1}^{H}\Phi_t^h(f)-\lambda\Delta f^1\left(x^1\right)+\ln\hat{p}_t(f)\right)+\alpha\eta\epsilon(2\beta+\epsilon)(t-1)H+\sum_{h=1}^{H}\kappa_1^h(\alpha,\epsilon)\\
&= \mathbb{E}_{S_{t-1}}\inf_p\mathbb{E}_{f\sim p}\left(\sum_{h=1}^{H}\Phi_t^h(f)-\lambda\Delta f^1\left(x^1\right)+\ln p(f)\right)+\alpha\eta\epsilon(2\beta+\epsilon)(t-1)H+\sum_{h=1}^{H}\kappa_1^h(\alpha,\epsilon)\\
&\leq \lambda\epsilon+\alpha\eta\epsilon(\epsilon+4\epsilon+2\beta)(t-1)H-\sum_{h=1}^{H}\ln p_0^h(\mathcal{F}(\epsilon,Q_{h+1}^*))+\sum_{h=1}^{H}\kappa_1^h(\alpha,\epsilon),
\end{aligned}$$

where the first inequality also uses the definition of $\Delta f_t^1(x^1)$; the second inequality comes from Lemma 10 and Lemma 11; the equality is because Lemma 5, and the last step comes from Lemma 8. Summing over t, we obtain that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\left(V_1^*(x^1)-V_1^{\mu_t,\nu_t}(x^1)\right)\\
&\leq \epsilon T+\frac{1}{\lambda}\alpha\eta(5\epsilon+2\beta)\frac{T(T-1)}{2}H-\frac{T}{\lambda}\sum_{h=1}^{H}\ln p_0^h(\mathcal{F}(\epsilon,Q_{h+1}^*))+\frac{T}{\lambda}\sum_{h=1}^{H}\kappa^h(\alpha,\epsilon)+\frac{\lambda}{\alpha\eta}dc(\mathcal{F},MG,T,0.25\alpha\eta/\lambda)\\
&\leq O(\beta\sqrt{dc(\mathcal{F},MG,T)\kappa(\beta/T^2)T}+dc(\mathcal{F},MG,T)).
\end{aligned}$$

Here in the last step, we first let $\alpha\to 1^-$ and note that

$$\begin{aligned}
-\ln p_0^h\left(\mathcal{F}(\epsilon,Q_{h+1}^*\right)&\leq\kappa_1^h(1,\epsilon),\\
-\sum_{h=1}^{H}\ln p_0^h\left(\mathcal{F}(\epsilon,Q_{h+1}^*\right)+\sum_{h=1}^{H}\kappa_1^h(1,\epsilon)&\leq 2\kappa(\epsilon).
\end{aligned}$$

Then, we take $\epsilon=\frac{\beta}{T^2}$, $\lambda=\sqrt{\frac{T\kappa(\beta/T^2)}{\beta^2 dc(\mathcal{F},MG,T)}}$, and $\eta=\frac{1}{4\beta^2}$. This concludes the proof. □

D. Proof of the Theorem 3

In this section, we provide a proof for Theorem 3.

Lemma 12. *For any max-player's policy μ and all functions $g\in\mathcal{F}$, we have*

$$\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)}\Delta L_\mu^h\left(g^h,g^{h+1},\zeta_s\right)=\left(\mathcal{E}_h^\mu\left(g;x_s^h,a_s^h,b_s^h\right)\right)^2$$

and

$$\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)}\Delta L_\mu^h\left(g^h,g^{h+1},\zeta_s\right)^2\leq\frac{4\beta^2}{3}\left(\mathcal{E}_h^\mu\left(g;x_s^h,a_s^h,b_s^h\right)\right)^2$$

Proof. The proof of this lemma only employs the Markov property of the transition and the range of function $g\in\mathcal{F}$. By replacing the notations in the proof of Lemma 6, we conclude the proof. □

Lemma 13. *Letting $\eta\beta^2\leq 0.8$, then for all functions $g\in\mathcal{F}$ and any max-player's policy μ, we have*

$$\begin{aligned}
&\ln\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)}\exp\left(-\eta\Delta L_\mu^h\left(g^h,g^{h+1},\zeta_s\right)\right)\\
&\quad\leq\mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)}\exp\left(-\eta\Delta L_\mu^h\left(g^h,g^{h+1},\zeta_s\right)\right)-1\\
&\quad\leq -0.25\eta\left(\mathcal{E}_h^\mu\left(f;x_s^h,a_s^h,b_s^h\right)\right)^2.
\end{aligned}$$

Proof. The proof of this lemma only employs the range of function $g \in \mathcal{F}$. By replacing the notations in the proof of Lemma 7, we conclude the proof. □

Lemma 14. *It holds that*

$$\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\inf_p \mathbb{E}_{g\sim p(\cdot)}\left[\sum_{h=1}^H \Phi_t^h(g,\mu_t) - \lambda\Delta g_{\mu_t}^1\left(x^1\right) + \ln p(g)\right]$$
$$\leq \lambda\epsilon + 4\alpha\eta(t-1)H\epsilon^2 - \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^H \ln p_0^h\left(\mathcal{F}_h^{\mu_t}\left(\epsilon, Q_{h+1}^{\mu_t,\dagger}\right)\right)$$

Proof. Consider any fixed $g \in \mathcal{F}$. For any $\tilde{g}^h \in \mathcal{F}^h$ that depends on S_{s-1} and μ_{f_s}, and for any μ_f we obtain from Lemma 7 that

$$\mathbb{E}_{\zeta_s}\exp\left(-\eta\Delta L_{\mu_f}^h\left(\tilde{g}^h, g^{h+1}, \zeta_s\right)\right) - 1 \leq -0.25\eta\mathbb{E}_{\zeta_s}\left(\tilde{g}^h(x,a) - \mathcal{T}_h^{\mu_f} g^{h+1}(x,a,b)\right)^2 \leq 0.$$

We now fix some t. For all $s \leq t$, we define

$$W_s^h := \mathbb{E}_{S_s}\mathbb{E}_{f\sim\hat{p}_{t+1}}\mathbb{E}_{g\sim p(\cdot)}\ln\mathbb{E}_{\tilde{g}^h\sim p_0^h}\exp\left(-\eta\sum_{\ell=1}^s \Delta L_{\mu_f}^h(\tilde{g}^h, g^{h+1}, \zeta_\ell)\right),$$

and recall that

$$\hat{q}_t^h\left(\tilde{g}^h \mid g^{h+1}, \mu_f, S_{t-1}\right) = \frac{p_0^h(\tilde{g}^h)\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_{\mu_f}^h\left(\tilde{g}^h, g^{h+1}, \zeta_s\right)\right)}{\mathbb{E}_{\tilde{g}'\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_{\mu_f}^h\left(\tilde{g}', g^{h+1}, \zeta_s\right)\right)}.$$

We have

$$W_s^h - W_{s-1}^h = \mathbb{E}_{S_s}\mathbb{E}_{f\sim\hat{p}_{t+1}(\cdot)}\ln\mathbb{E}_{\tilde{g}^h\sim\hat{q}_s^h(\cdot|g^{h+1},\mu_f,S_{s-1})}\exp\left(-\eta\Delta L_{\mu_f}^h\left(\tilde{g}^h, g^{h+1}, \zeta_s\right)\right)$$
$$\leq \mathbb{E}_{S_s}\mathbb{E}_{f\sim\hat{p}_t(\cdot)}\left(\mathbb{E}_{\tilde{g}^h\sim\hat{q}_s^h(\cdot|g^{h+1},\mu_f,S_{s-1})}\exp\left(-\eta\Delta L_{\mu_f}^h\left(\tilde{g}^h, g^{h+1}, \zeta_s\right)\right) - 1\right) \leq 0$$

where we use $\ln z \leq z - 1$. By $W_0^h = 0$, we know that

$$W_t^h = W_0^h + \sum_{s=1}^t [W_s^h - W_{s-1}^h] \leq 0,$$

equivalently,

$$\mathbb{E}_{S_t}\mathbb{E}_{f\sim\hat{p}_{t+1}(\cdot)}\mathbb{E}_{g\sim p(\cdot)}\ln\mathbb{E}_{\tilde{g}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^t \Delta L_{\mu_f}^h(\tilde{g}^h, g^{h+1}, \zeta_s)\right) \leq 0.$$

Note that t is arbitrary. This implies that for any $p(\cdot)$ and any t, we have

$$\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t(\cdot)}\mathbb{E}_{g\sim p(\cdot)}\left[\sum_{h=1}^H \Phi_t^h(g,\mu_f) - \lambda\Delta g_{\mu_f}^1\left(x^1\right) + \ln p(g)\right]$$
$$= \mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t(\cdot)}\mathbb{E}_{g\sim p(\cdot)}\left[-\lambda\Delta g_{\mu_f}^1\left(x^1\right) + \alpha\eta\sum_{h=1}^H\sum_{s=1}^{t-1}\Delta L_{\mu_f}^h\left(g^h, g^{h+1}, \zeta_s\right)\right.$$
$$\left. +\alpha\sum_{h=1}^H \ln\mathbb{E}_{\tilde{g}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_{\mu_f}^h\left(\tilde{g}^h, g^{h+1}, \zeta_s\right)\right) + \ln\frac{p(g)}{p_0(g)}\right]$$
$$\leq \mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t(\cdot)}\mathbb{E}_{g\sim p(\cdot)}\left[-\lambda\Delta g_{\mu_f}^1\left(x^1\right) + \sum_{h=1}^H \alpha\eta\sum_{s=1}^{t-1}\left(\mathcal{E}_h^{\mu_f}\left(g; x_s^h, a_s^h, b_s^h\right)\right)^2 + \ln\frac{p(g)}{p_0(g)}\right].$$

Since $p(\cdot)$ is arbitrary, we can take $g^h \in \mathcal{F}_h^\mu(\epsilon, Q_{h+1}^{\mu,\dagger})$ for all $h \in [H]$. We need to show that g^h admits a small $\mathcal{T}_h^\mu$-Bellman-residual. We have

$$|g^h(x,a,b) - Q_h^{\mu,\dagger}(x,a,b)| = |g^h(x,a,b) - \mathcal{T}_h^\mu Q_{h+1}^{\mu,\dagger}(x,a,b)| \leq \epsilon,$$

for all $(x,a,b,h) \in \mathcal{X}\times\mathcal{A}\times\mathcal{B}\times[H]$. Then, we have

$$|\mathcal{E}_h^{\mu}(g;x,a,b)| \le |g^h(x,a,b) - Q_h^{\mu,\dagger}(x,a,b)| + \sup_{x'} |V_{g,h+1}^{\mu}(x') - V_{h+1}^{\mu,\dagger}(x')| \le 2\epsilon,$$

where we use

$$\begin{aligned}|V_{g,h+1}^{\mu}(x') - V_{h+1}^{\mu,\dagger}(x')| &= |\inf_{\nu} \mathbb{D}_{\mu,\nu} g(x') - \inf_{\nu} \mathbb{D}_{\mu,\nu} Q_{h+1}^{\mu,\dagger}(x')| \\ &\le \sup_{\nu} |\mathbb{D}_{\mu,\nu}(g(x') - Q_{h+1}^{\mu,\dagger}(x'))| \le \epsilon,\end{aligned}$$

where we use the fact that

$$|\inf_A f - \inf_A g| \le \sup_A |f-g|.$$

By taking $p(f) = p_0(f) I(f \in \mathcal{F}(\epsilon,\mu_f))/p_0(\mathcal{F}(\epsilon,\mu_f))$, with $\mathcal{F}(\epsilon,\mu_f) = \prod_h \mathcal{F}_h^{\mu_f}(\epsilon, Q_{h+1}^{\mu_f,\dagger})$, we obtain the desired result. □

Lemma 15. *For any max-player's policy μ that is induced by some $f \in \mathcal{F}$, we have*

$$\begin{aligned}\mathbb{E}_{g\sim\hat{p}_t^{\mu}(g)} \ln \hat{p}_t^{\mu}(g) \ge& \alpha \mathbb{E}_{g\sim\hat{p}_t^{\mu}} \ln \hat{p}_t^{\mu}(g) + (1-\alpha)\mathbb{E}_{g\sim\hat{p}_t^{\mu}} \sum_{h=1}^{H} \ln \hat{p}_t^{\mu}\left(g^h\right) \\ \ge& \frac{\alpha}{2}\sum_{h=1}^{H} \mathbb{E}_{g\sim\hat{p}_t^{\mu}} \ln \hat{p}_t^{\mu}\left(g^h, g^{h+1}\right) \\ &+ (1-0.5\alpha)\mathbb{E}_{g\sim\hat{p}_t^{\mu}} \ln \hat{p}_t^{\mu}\left(g^1\right) + (1-\alpha)\sum_{h=2}^{H} \mathbb{E}_{g\sim\hat{p}_t^{\mu}} \ln \hat{p}_t^{\mu}\left(g^h\right).\end{aligned} \tag{D.1}$$

Proof. The proof of this lemma only relies on the non-negativity of mutual information and KL-divergence. By replacing the notations in the proof of Lemma 9, we conclude the proof. □

Lemma 16. *It holds that*

$$\begin{aligned}&\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu}}\left(\sum_{h=1}^{H}\Phi_t^h(g) - \lambda\Delta g_{\mu_t}^1\left(x^1\right) + \ln\hat{p}_t^{\mu_t}(g)\right) \\ &\ge \underbrace{\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu}}\left[-\lambda\Delta g_{\mu_t}^1\left(x^1\right) + (1-0.5\alpha)\ln\frac{\hat{p}_t^{\mu_t}\left(g^1\right)}{p_0^1\left(g^1\right)}\right]}_{A'} \\ &\quad + \sum_{h=1}^{H}\underbrace{0.5\alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\left[\eta\sum_{s=1}^{t-1}2\Delta L_{\mu_t}^h\left(g^h,g^{h+1},\zeta_s\right) + \ln\frac{\hat{p}_t^{\mu_t}\left(g^h,g^{h+1}\right)}{p_0^h\left(g^h\right)p_0^{h+1}\left(g^{h+1}\right)}\right]}_{B_h'} \\ &\quad + \sum_{h=1}^{H}\underbrace{\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\left[\alpha\ln\mathbb{E}_{\tilde{g}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_{\mu_t}^h\left(\tilde{g}^h,g^{h+1},\zeta_s\right)\right) + (1-\alpha)\ln\frac{\hat{p}_t^{\mu_t}\left(g^{h+1}\right)}{p_0^{h+1}\left(g^{h+1}\right)}\right]}_{C_h'}.\end{aligned} \tag{D.2}$$

Proof. We use the definition of the potential function and apply Lemma 15 (note that it is valid for any $\mu_f, f\in\mathcal{F}$). □

Lemma 17. *If $\eta\beta^2 \le 0.4$, it holds that*

$$A' \ge -\lambda\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\Delta g_{\mu_t}^1(x^1), \tag{D.3}$$

$$B_h' \ge 0.25\alpha\eta\sum_{s=1}^{t-1}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\mathbb{E}_{\pi_s}\left(\mathcal{E}_h^{\mu_t}\left(f;x_s^h,a_s^h,b_s^h\right)\right)^2 \tag{D.4}$$

$$C_h' \ge -\alpha\eta\epsilon(2b+\epsilon)(t-1) - \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\kappa_{\mu_t}^h(\alpha,\epsilon). \tag{D.5}$$

Proof. The bound of A' comes from the non-negativity of KL-divergence and $\alpha \in (0,1]$. To prove the lower bound of B'_h, we define

$$\xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right) = -2\eta\Delta L_{\mu_t}^h\left(g^h, g^{h+1}, \zeta_s\right) - \ln \mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)} \exp\left(-2\eta\Delta L_{\mu_t}^h\left(g^h, g^{h+1}, \zeta_s\right)\right),$$

where μ_t is an arbitrary policy induced by some $f_t \in \mathcal{F}$. Then, for all $h \in [H]$, we have

$$\mathbb{E}_{S_{t-1}} \exp\left(\sum_{s=1}^{t-1} \xi_s^h\left(g^h, g^{h+1}, \zeta_s\right)\right) = 1,$$

according to Lemma 3. Then, by Lemma 4, we have

$$\begin{aligned}
&\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\left[\sum_{s=1}^{t-1} -\xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right) + \ln\frac{\hat{p}_t^{\mu_t}\left(g^h, g^{h+1}\right)}{p_0^h\left(g^h\right)p_0^{h+1}\left(g^{h+1}\right)}\right] \\
&\quad\geq \inf_p \mathbb{E}_{g\sim p}\left[\sum_{s=1}^{t-1} -\xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right) + \ln\frac{p\left(g^h, g^{h+1}\right)}{p_0^h\left(g^h\right)p_0^{h+1}\left(g^{h+1}\right)}\right] \\
&\quad= -\ln \mathbb{E}_{g^h\sim p_0^h}\mathbb{E}_{g^{h+1}\sim p_0^{h+1}} \exp\left(\sum_{s=1}^{t-1} \xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right)\right),
\end{aligned}$$

where the last step is from some simple calculations and the fact that Lemma 4 implies that the $\inf$ is achieved by $p(g^h, g^{h+1}) \propto p_0^h(g^h)p_0^{h+1}(g^{h+1})\exp(\sum_{s=1}^{t-1}\xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right))$.

This implies that

$$\begin{aligned}
&\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\left[\sum_{s=1}^{t-1} -\xi_s^h\left(f^h, f^{h+1}, \mu_t, \zeta_s\right) + \ln\frac{\hat{p}_t^{\mu_t}\left(f^h, f^{h+1}\right)}{p_0^h\left(f^h\right)p_0^{h+1}\left(f^{h+1}\right)}\right] \\
&\quad\geq -\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\ln \mathbb{E}_{g^h\sim p_0^h}\mathbb{E}_{g^{h+1}\sim p_0^{h+1}} \exp\left(\sum_{s=1}^{t-1} \xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right)\right) \\
&\quad\geq -\ln \mathbb{E}_{g^h\sim p_0^h}\mathbb{E}_{g^{h+1}\sim p_0^{h+1}}\mathbb{E}_{f\sim\hat{p}_t}\mathbb{E}_{S_{t-1}} \exp\left(\sum_{s=1}^{t-1} \xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right)\right) = 0,
\end{aligned}$$

where we use the convexity of $-\ln(\cdot)$. With this result, the definition of B'_h and the definition of $\xi_s^h\left(g^h, g^{h+1}, \mu_t, \zeta_s\right)$, we have

$$\begin{aligned}
B'_h &= 0.5\alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\left[\eta\sum_{s=1}^{t-1} 2\Delta L_{\mu_t}^h\left(g^h, g^{h+1}, \zeta_s\right) + \ln\frac{\hat{p}_t^{\mu_t}\left(g^h, g^{h+1}\right)}{p_0^h\left(g^h\right)p_0^{h+1}\left(g^{h+1}\right)}\right] \\
&\geq 0.5\alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\sum_{s=1}^{t-1} -\ln \mathbb{E}_{x_s^{h+1}\sim\mathbb{P}^h(\cdot|x_s^h,a_s^h,b_s^h)} \exp\left(-2\eta\Delta L_{\mu_t}^h\left(g^h, g^{h+1}, \zeta_s\right)\right) \\
&\geq -0.5\alpha\eta\sum_{s=1}^{t-1}\frac{1}{2}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\mathbb{E}_{g\sim\hat{p}_t^{\mu_t}}\mathbb{E}_{\pi_s}(\mathcal{E}_h^{\mu_t}(g; x_s^h, a_s^h, r_s^h))^2,
\end{aligned}$$

where we use Lemma 13 in the last step.

We now turn to the lower bound of C_h. For any max-player's policy μ, we have

$$\begin{aligned}
&\mathbb{E}_{g\sim\hat{p}_t^{\mu}}\left[\alpha\ln \mathbb{E}_{\tilde{g}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_{\mu}^h\left(\tilde{g}^h, g^{h+1}, \zeta_s\right)\right) + (1-\alpha)\ln\frac{\hat{p}_t^{\mu}\left(g^{h+1}\right)}{p_0^{h+1}\left(g^{h+1}\right)}\right] \\
\geq&(1-\alpha)\inf_{p^h}\mathbb{E}_{g\sim p^h}\left[\frac{\alpha}{1-\alpha}\ln \mathbb{E}_{\tilde{g}^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_{\mu}^h\left(\tilde{g}^h, g^{h+1}, \zeta_s\right)\right) + \ln\frac{p^h\left(g^{h+1}\right)}{p_0^{h+1}\left(g^{h+1}\right)}\right] \\
=&-(1-\alpha)\ln \mathbb{E}_{g^{h+1}\sim p_0^{h+1}}\left(\mathbb{E}_{g^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_{\mu}^h\left(g^h, g^{h+1}, \zeta_s\right)\right)\right)^{-\alpha/(1-\alpha)},
\end{aligned}$$

where we use the fact that the $\inf$ is achieved at

$$p^h\left(g^{h+1}\right) \propto p_0^{h+1}\left(g^{h+1}\right)\left(\mathbb{E}_{g^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_\mu^h\left(g^h,g^{h+1},\zeta_s\right)\right)\right)^{-\alpha/(1-\alpha)}.$$

We now consider a fixed $g^h\in\mathcal{F}_h^\mu(\epsilon,g^{h+1})$. Using the same arguments as in the proof of Lemma 11, it holds that

$$\left|\Delta L_\mu^h\left(g^h,g^{h+1},\zeta_s\right)\right| \le \left(\mathcal{E}_h^\mu\left(g,x_s^h,a_s^h\right)\right)^2 + 2b\left|\mathcal{E}_h^\mu\left(g,x_s^h,a_s^h\right)\right| \le \epsilon(2b+\epsilon).$$

It follows that

$$\mathbb{E}_{g^h\sim p_0^h}\exp\left(-\eta\sum_{s=1}^{t-1}\Delta L_\mu^h\left(g^h,g^{h+1},\zeta_s\right)\right) \le p_0^h\left(\mathcal{F}_h^\mu\left(\epsilon,g^{h+1}\right)\right)\exp(-\eta(t-1)(2b+\epsilon)\epsilon)$$

Thus, we have

$$\begin{aligned}
C_h &\ge \alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\ln\mathbb{E}_{f^{h+1}\sim p_0^{h+1}}p_0^h(\mathcal{F}_h^{\mu_t}(\epsilon,g^{h+1}))\exp(-\eta(t-1)(2b+\epsilon)\epsilon)\\
&= -\alpha\eta\epsilon(2b+\epsilon)(t-1) + \alpha\mathbb{E}_{S_{t-1}}\mathbb{E}_{f\sim\hat{p}_t}\ln\mathbb{E}_{g^{h+1}\sim p_0^{h+1}}p_0^h(\mathcal{F}_h^{\mu_t}(\epsilon,g^{h+1}))\\
&\ge -\alpha\eta\epsilon(2b+\epsilon)(t-1) - \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\kappa_{\mu_t}^h(\alpha,\epsilon),
\end{aligned}$$

where we use the definition

$$\kappa_\mu^h(\alpha,\epsilon) = (1-\alpha)\ln\mathbb{E}_{g^{h+1}\sim p_0^{h+1}}p_0^h\left(\mathcal{F}_h^\mu\left(\epsilon,g^{h+1}\right)\right)^{-\alpha/(1-\alpha)}.$$

□

We are ready to prove Theorem 3.

Proof of Theorem 3. Let π_t denote the distribution induced by $\mu_t\times\nu_t$ and define

$$\delta_t^h = -\lambda\mathcal{E}_h^{\mu_t}(g_t,x_t^h,a_t^h,b_t^h) - 0.25\alpha\eta\sum_{s=1}^{t-1}\mathbb{E}_{\pi_s}\left(\mathcal{E}_h^{\mu_t}(g_t,x_t^h,a_t^h,b_t^h)\right)^2.$$

According to the value-decomposition Lemma 2, we have

$$\begin{aligned}
&\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\lambda(V_1^{\mu_t,\nu_t}(x^1) - V_1^{\mu_t,\dagger}(x^1)) - \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\mathbb{E}_{\zeta_t\sim\pi_t}\sum_{h=1}^H\delta_t^h\\
&= -\lambda\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\Delta g_{t,1}^{\mu_t}(x^1) + 0.25\alpha\eta\sum_{h=1}^H\sum_{s=1}^{t-1}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t^{\mu_t}}\mathbb{E}_{\pi_s}\left(\mathcal{E}_h^{\mu_t}(g_t,x_t^h,a_t^h,b_t^h)\right)^2\\
&\le \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t}\left(\sum_{h=1}^H\Phi_t^h(g_t,\mu_t) - \lambda\Delta g_{t,1}^{\mu_t}(x^1) + \ln\hat{p}_t^{\mu_t}(g_t)\right)\\
&\qquad + \alpha\eta\epsilon(2b+\epsilon)(t-1)H + \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^H\kappa_{\mu_t}^h(\alpha,\epsilon)\\
&= \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\inf_p\mathbb{E}_{g\sim p}\left(\sum_{h=1}^H\Phi_t^h(g,\mu_t) - \lambda\Delta g_{t,1}^{\mu_t}(x^1) + \ln p(g)\right)\\
&\qquad + \alpha\eta\epsilon(2\beta+\epsilon)(t-1)H + \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^H\kappa_{\mu_t}^h(\alpha,\epsilon)\\
&\le \lambda\epsilon + \alpha\eta\epsilon(\epsilon+4\epsilon+2\beta)(t-1)H - \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^H\ln p_0^h(\mathcal{F}_h(\epsilon,Q_{h+1}^{\mu_t,\dagger})) + \mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^H\kappa_{\mu_t}^h(\alpha,\epsilon).
\end{aligned}$$

Summing over t, we obtain that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\mathbb{E}_{g_t\sim\hat{p}_t}(V_1^{\mu_t,\nu_t}(x^1)-V_1^{\mu_t,\dagger}(x^1))\\
&\leq \epsilon T+\frac{1}{\lambda}\alpha\eta(5\epsilon+2\beta)\frac{T(T-1)}{2}H-\frac{1}{\lambda}\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^{H}\ln p_0^h(\mathcal{F}_h(\epsilon,Q_{h+1}^{\mu_t,\dagger}))\\
&+\frac{1}{\lambda}\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^{H}\kappa_{\mu_t}^h(\alpha,\epsilon)+\frac{\lambda}{\alpha\eta}dc(\mathcal{F},MG,T,0.25\alpha\eta/\lambda)\\
&\leq O(\beta\sqrt{dc(\mathcal{F},MG,T)\kappa(\beta/T^2)T}+dc(\mathcal{F},MG,T)).
\end{aligned}$$

The last step is proved as follows. We find an upper bound for $\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{t=1}^{T}\sum_{h=1}^{H}\kappa_{\mu_t}^h(\alpha,\epsilon)$. We note that for all μ_t, $\kappa_{\mu_t}^h(\alpha,\epsilon)$ is increasing w.r.t. α with the limit $\kappa_{\mu_t}^h(1,\epsilon)\leq\kappa(\epsilon)<\infty$. By monotone convergence theorem, we know that

$$\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{t=1}^{T}\sum_{h=1}^{H}\kappa_{\mu_t}^h(\alpha,\epsilon)\to\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{t=1}^{T}\sum_{h=1}^{H}\kappa_{\mu_t}^h(1,\epsilon)=\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\kappa_{\mu_t}(\epsilon)\leq T\kappa(\epsilon).$$

We also have

$$\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{t=1}^{T}\sum_{h=1}^{H}-\ln p_0^h\left(\mathcal{F}_h(\epsilon,Q_{h+1}^{\mu_t,\dagger})\right)\leq\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{t=1}^{T}\kappa_{\mu_t}(\epsilon)\leq T\kappa(\epsilon).$$

It follows that

$$-\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^{H}\ln p_0^h(\mathcal{F}_h(\epsilon,Q_{h+1}^{\mu_t,\dagger}))+\sum_{t=1}^{T}\mathbb{E}_{S_{t-1}}\mathbb{E}_{f_t\sim\hat{p}_t}\sum_{h=1}^{H}\kappa_{\mu_t}^h(1,\epsilon)\leq 2T\kappa(\epsilon).$$

Now we first let $\alpha\to 1^-$. Then, we take $\epsilon=\frac{\beta}{T^2}$, $\lambda=\sqrt{\frac{T\kappa(\beta/T^2)}{\beta^2 dc(\mathcal{F},MG,T)}}$, $\eta=\frac{1}{4\beta^2}$. This concludes the proof. □

E. Proof of the Value-Decomposition Lemma

Proof of Lemma 1. Let $\mu=\mu_f$ and ν be an arbitrary policy taken by the min-player.

$$\begin{aligned}
&V_1^*(x^1)-V_1^{\mu,\nu}(x^1)\\
&\quad=\sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}V_{f,h}(x^h)-r^h(x^h,a^h,b^h)-V_{f,h+1}(x^{h+1})+V_1^*(x^1)-V_{f,1}(x^1)\\
&\quad=\sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}\min_{\nu'}\mathbb{D}_{\mu,\nu'}f(x^h)-r^h(x^h,a^h,b^h)-V_{f,h+1}(x^{h+1})+V_1^*(x^1)-V_{f,1}(x^1)\\
&\quad\leq\sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}\mathbb{D}_{\mu,\nu}f^h(x^h)-r^h(x^h,a^h,b^h)-V_{f,h+1}(x^{h+1})+V_1^*(x^1)-V_{f,1}(x^1)\\
&\quad=\sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}f^h(x^h,a^h,b^h)-r^h(x^h,a^h,b^h)-V_{f,h+1}(x^{h+1})+V_1^*(x^1)-V_{f,1}(x^1)\\
&\quad=\sum_{h=1}^{H}\mathbb{E}_{\mu,\nu}\mathcal{E}_h(f^h,f^{h+1},\zeta)+V_1^*(x^1)-V_{f,1}(x^1),
\end{aligned}$$

where the first equality comes from the value-decomposition Theorem (Jiang et al., 2017) (can be verified easily by telescope sum and $V^{H+1}=0$); the second equality is because of the definition of $\mu=\mu_{f,h}(x)=\operatorname{argmax}_{\mu\in\Delta_{\mathcal{A}}}\min_{\nu\in\Delta_{\mathcal{B}}}\mu^\top f^h(x,\cdot,\cdot)\nu$; the inequality comes from the fact that $\mu=\mu_f$ and ν may not be $\operatorname{argmin}_{\nu'}\mathbb{D}_{\mu,\nu'}f(x^h)$. This decomposition accounts for the use of an optimistic prior in Algorithm 2. □

Proof of Lemma 2. Suppose that $\mu = \mu_f$ is taken by the max-player and g is sampled from the posterior by the booster agent. Let ν be given by $\nu = \operatorname{argmin}_{\nu'} V_h^{\mu}(x)$ for all (x, h). Then, we have:

$$\begin{aligned}
&V_1^{\mu,\dagger}(x^1) - V_1^{\mu,\nu}(x^1) \\
&\quad = V_{g,1}^{\mu}(x^1) - V_1^{\mu,\nu}(x^1) + V_1^{\mu_t,\dagger}(x^1) - V_{g,1}^{\mu}(x^1) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu}\mathbb{D}_{\mu,\nu} g(x^h) - r^h(x^h, a^h, b^h) - V_{g,h+1}^{\mu}(x^{h+1}) + V_1^{\mu,\dagger}(x^1) - V_{g,1}^{\mu}(x^1) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu} g^h(x^h, a^h, b^h) - r^h(x^h, a^h, b^h) - V_{g,h+1}^{\mu}(x^{h+1}) + V_1^{\mu,\dagger}(x^1) - V_{g,1}^{\mu}(x^1) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{\mu,\nu} \mathcal{E}_h^{\mu}(g^h, g^{h+1}, \zeta) + V_1^{\mu,\dagger}(x^1) - V_{g,1}^{\mu}(x^1).
\end{aligned}$$

This decomposition accounts for the use of an optimistic prior in Algorithm 3. □

F. Proof of the Decoupling Coefficient Bounds

In this section, we provide proofs for the decoupling coefficient bounds. We need the following lemma.

Lemma 18 (Elliptical Potential Lemma, Lemma 10 of Xie et al. (2020))**.** *Suppose $\{\phi_t\}_{t\geq 0}$ is a sequence in $\mathbb{R}^d$ satisfying $\|\phi_t\| \leq 1$. Let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix, and $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t} \phi_i \phi_i^\top$. If the smallest eigenvalues of Λ_0 is lower bounded by 1, then*

$$\log\left(\frac{\det \Lambda_t}{\det \Lambda_0}\right) \leq \sum_{i\in[t]} \phi_i^\top \Lambda_{j-1}^{-1} \phi_i \leq 2\log\left(\frac{\det \Lambda_t}{\det \Lambda_0}\right).$$

Proof of Proposition 1. We first note that the completeness assumption is satisfied in linear MG case whose proof can be found in Huang et al. (2021). Now we consider two arbitrary θ^h, θ^{h+1} whose norms are bounded by $H\sqrt{d}$ and $f \in \mathcal{F}$. We also define a function $g \in \mathcal{F}$ s.t. $g^h = g(\theta^h)$ and $g^{h+1} = g(\theta^{h+1})$. By Assumption 2, we can find some $\theta^h(f) \in \mathbb{R}^d$ with $\left\|\theta^h(f)\right\| \leq H\sqrt{d}$ s.t. $\mathcal{T}_h^{\mu_f}(\phi(x,a,b)^\top \theta^{h+1}) = \phi(x,a,b)^\top \theta^h(f)$. Therefore, we have

$$\mathcal{E}_h^{\mu_f}(g; x, a, b) = \phi(x,a,b)^\top(\theta^h - \theta^h(f)) = \phi(x,a,b)^\top w^h(f,g),$$

where $w^h(f,g) \in \mathbb{R}^d$ satisfies $\left\|w^h(f,g)\right\| \leq 2H\sqrt{d}$. We denote $\phi_s^h = \mathbb{E}_{\pi_s}[\phi(x^h, a^h, b^h)]$ and $\Phi_t^h = \lambda I + \sum_{s=1}^{t} \phi(x^h, a^h, b^h)\phi(x^h, a^h, b^h)^\top$ where $\lambda \geq 1$ is a tuning parameter. Then, we have

$$\begin{aligned}
&\mathbb{E}_{\pi_t}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_t^h, a_t^h, b_t^h\right)\right] - \mu \sum_{s=1}^{t-1} \mathbb{E}_{\pi_s}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_s^h, a_s^h\right)^2\right] \\
&= w^h\left(f_t, g_t\right)^\top \phi_t^h - \mu w^h\left(f_t, g_t\right)^\top \sum_{s=1}^{t-1} \mathbb{E}_{\pi_s}\left[\phi\left(x^h, a^h, b^h\right)\phi\left(x^h, a^h, b^h\right)^\top\right] w^h\left(f_t, g_t\right) \\
&\leq w^h\left(f_t, g_t\right)^\top \phi_t^h - \mu w^h\left(f_t, g_t\right)^\top \Phi_{t-1}^h w^h\left(f_t, g_t\right) + 4\mu\lambda d H^2 \\
&\leq \frac{1}{4\mu}\left(\phi_t^h\right)^\top \left(\Phi_{t-1}^h\right)^{-1} \phi_t^h + 4\mu\lambda d H^2
\end{aligned}$$

where the first inequality uses Jensen's inequality and $\left\|w^h(f_t, g_t)\right\| \leq 2H\sqrt{d}$ and the second inequality comes from the

fact $(a^\top b) \leq (\|a\|_{\Phi_{t-1}^h} \|b\|_{(\Phi_{t-1}^h)^{-1}}) \leq \frac{1}{2}(\|a\|^2_{\Phi_{t-1}^h} + \|b\|^2_{(\Phi_{t-1}^h)^{-1}})$. Summing over $t \in [T]$ and $h \in [H]$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_t}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_t^h, a_t^h, b_t^h\right)\right] - \mu \sum_{s=1}^{t-1} \mathbb{E}_{\pi_s}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_s^h, a_s^h\right)^2\right] \\
&\leq \sum_{h=1}^{H}\left[\frac{\ln(\det(\Phi_T^h) - d\ln(\lambda)}{2\mu} + 4\mu\lambda d H^2 T\right] \\
&\leq H(\frac{d\ln(\lambda + T/d) - d\ln(\lambda)}{2\mu} + 4\mu\lambda d H^2 T),
\end{aligned}$$

where we use the Elliptical Potential lemma in the first inequality and the second inequality uses

$$\ln\det(\Phi_T^h) \leq d\ln\frac{\text{trace}(\Phi_T^h)}{d}, \text{ and, } \text{trace}(\Phi_t^h) \leq \lambda d + T.$$

By setting $\lambda = \min\{1, \frac{1}{\mu^2 H^2 T}\}$, we conclude the proof. □

Proof of Proposition 2. We assume that $c_1 \leq 1 \leq c_2$. Otherwise, we can scale the feature maps and the link function accordingly. By similar arguments with the completeness assumption as in the proof of Proposition 1, we have

$$\mathcal{E}_h^{\mu_f}(g; x, a, b) = \sigma(\phi(x, a, b)^\top \theta^h) - \sigma(\phi(x, a, b)^\top \theta^h(f)).$$

By the Lipschitz property, we have

$$c_1 |\phi(x, a, b)^\top w(f, g)| \leq |\mathcal{E}_h^{\mu_f}(g; x, a, b)| \leq c_2 |\phi(x, a, b)^\top w(f, g)|,$$

for some $w(f, g) \in \mathbb{R}^d$ satisfying $w(f, g) \leq 2H\sqrt{d}$. We denote $\phi_h^s = \mathbb{E}_{\pi_s}[\phi(x^h, a^h, b^h)]$ and $\Phi_t^h = \lambda I + \sum_{s=1}^{t} \phi(x^h, a^h, b^h)\phi(x^h, a^h, b^h)^\top$ where $\lambda \geq 1$ is a tuning parameter. Then, we have

$$\begin{aligned}
&\mathbb{E}_{\pi_t}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_t^h, a_t^h, b_t^h\right)\right] - \mu \sum_{s=1}^{t-1} \mathbb{E}_{\pi_s}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_s^h, a_s^h\right)^2\right] \\
&\leq c_2 |w^h\left(f_t, g_t\right)^\top \phi_t^h| - \mu c_1^2 w^h\left(f_t, g_t\right)^\top \sum_{s=1}^{t-1} \mathbb{E}_{\pi_s}\left[\phi\left(x^h, a^h, b^h\right)\phi\left(x^h, a^h, b^h\right)^\top\right] w^h\left(f_t, g_t\right) \\
&\leq c_2 |w^h\left(f_t, g_t\right)^\top \phi_t^h| - \mu c_1^2 w^h\left(f_t, g_t\right)^\top \Phi_{t-1}^h w^h\left(f_t, g_t\right) + 4\mu c_1^2 \lambda d H^2 \\
&\leq \frac{c_2^2}{4\mu c_1^2}\left(\phi_t^h\right)^\top \left(\Phi_{t-1}^h\right)^{-1} \phi_t^h + 4\mu c_1^2 \lambda d H^2.
\end{aligned}$$

Summing over $t \in [T]$ and $h \in [H]$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_t}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_t^h, a_t^h, b_t^h\right)\right] - \mu \sum_{s=1}^{t-1} \mathbb{E}_{\pi_s}\left[\mathcal{E}_h^{\mu_{f_t}}\left(g_t; x_s^h, a_s^h\right)^2\right] \\
&\leq \sum_{h=1}^{H} c_2^2\left[\frac{\ln(\det(\Phi_T^h) - d\ln(\lambda)}{2\mu c_1^2} + 4\mu\lambda c_1^2 H^2 d T\right] \\
&\leq H c_2^2(\frac{d\ln(\lambda + T/d) - d\ln(\lambda)}{2\mu c_1^2} + 4\mu\lambda c_1^2 H^2 d T).
\end{aligned}$$

Setting $\lambda = \min\{1, \frac{1}{\mu^2 c_1^2 H^2 T}\}$ concludes the proof. □

In what follows, we prove the reduction of Bellman-Eluder dimension to the decoupling coefficient. The proof is almost the same as that of Dann et al. (2021) with minor modification. We start with the following lemma from (Dann et al., 2021).

Lemma 19. *Let $\pi_1, \cdots, \pi_{t-1}$ denote the measures over $\mathcal{X}\times\mathcal{A}\times\mathcal{B}$ obtained by following the policy induced by $(f_s, g_s)_{s=1}^{t-1} \in \mathcal{F}\times\mathcal{F}$ at stage h and $\{d_1, \cdots, d_M\}$ be the set of unique measures in decreasing order of occurrences and let $(N_i)_{i=1}^M$ be the number of times a measure appears in the sequence. If the ϵ-BE dimension for the distribution set $\mathcal{D}_{\mathcal{F}}$ and the function classes:*

$$\{g^h - \mathcal{T}_h^{\mu_f} g^{h+1} : f, g \in \mathcal{F}\}.$$

is E and $|\mathcal{E}_{x,a,b\sim\pi_t}\mathcal{E}_h^{\mu_{f_t}}(g_t; x, a, b)| > \epsilon$. Then, it holds that

$$\begin{aligned}&\sum_{s=1}^{t-1}\mathbb{E}_{x,a,b\sim\pi_s}\left[\mathcal{E}_h^{\mu_{f_t}}(g_t; x, a, b)^2\right] \geq w_t^h\left(\mathbb{E}_{x,a,b\sim\pi_t}\left[\mathcal{E}_h^{\mu_{f_t}}(g_t; x, a, b)\right]\right)^2\\ &\text{where } w_t^h = \begin{cases} N_i & \text{if } \pi_t = d_i \wedge i \in [E-1]\\ \left\lceil \frac{\sum_{i=E}^M N_i}{E}\right\rceil & \text{otherwise.}\end{cases}\end{aligned} \tag{F.1}$$

Proof of Proposition 3. Denote $\epsilon_{t,s}^h = \mathbb{E}_{x_s^h, a_s^h, b_s^h}\mathcal{E}_h^{\mu_{f_t}}(g_t; x, a, b)$, the LHS is upper bounded by

$$\sum_{t=1}^T\sum_{h=1}^H \epsilon_{tt}^h \leq EH + \epsilon TH + \sum_{t=E+1}^T\sum_{h=1}^H \epsilon_{tt}^h \mathbb{I}\left\{\epsilon_{tt}^h > \epsilon\right\}.$$

For each $h \in [H]$, the RHS is bounded by

$$\begin{aligned}\mu\sum_{t=1}^T\sum_{s=1}^{t-1}{\epsilon_{ts}^h}^2 + \frac{2E(1+\ln(T))}{4\mu} &\geq \sqrt{2E(1+\ln(T))}\sum_{t=E+1}^T w_t^h {\epsilon_{tt}^h}^2 \mathbb{I}\left\{\epsilon_{tt}^h > \epsilon\right\}\\ &\geq \sqrt{\frac{2E(1+\ln(T))}{\sum_{t=E+1}^T \frac{1}{w_t^h}}\sum_{t=E+1}^T \epsilon_{tt}^h \mathbb{I}\left\{\epsilon_{tt}^h > \epsilon\right\}},\end{aligned}$$

where the last inequality is because

$$\sqrt{\sum_t \frac{1}{w_t^h}\sum_t w_t^h {\epsilon_{tt}^h}^2 \mathbb{I}^2\left\{\epsilon_{tt}^h > \epsilon\right\}} \geq \sqrt{\sum_t \epsilon_{tt}^h \mathbb{I}\left\{\epsilon_{tt}^h > \epsilon\right\}},$$

which is due to

$$\sqrt{\sum x_i y_i} \leq \sqrt{\sum x_i^2 \sum y_i^2}.$$

Every time a measure π_t appears in the $E-1$ most common measures, one of N_i increases. Otherwise, $\sum_{i\geq E} N_i$ increases by 1. Hence,

$$\sum_{t=1}^T \frac{1}{w_t^h} \leq \sum_{i=1}^{E-1}\sum_{t=1}^T \frac{1}{t} + \sum_{t=1}^T \frac{E}{t} \leq 2E(1+\ln(T)).$$

□