Pessimistic Minimax Value Iteration: Provably Efficient Equilibrium Learning from Offline Datasets

Han Zhong [*1] **Wei Xiong** [*2] **Jiyuan Tan** [*3] **Liwei Wang** [1 4] **Tong Zhang** [2 5] **Zhaoran Wang** [6] **Zhuoran Yang** [7]

Abstract

We study episodic two-player zero-sum Markov games (MGs) in the offline setting, where the goal is to find an approximate Nash equilibrium (NE) policy pair based on a dataset collected a priori. When the dataset does not have uniform coverage over all policy pairs, finding an approximate NE involves challenges in three aspects: (i) distributional shift between the behavior policy and the optimal policy, (ii) function approximation to handle large state space, and (iii) minimax optimization for equilibrium solving. We propose a pessimism-based algorithm, dubbed as pessimistic minimax value iteration (PMVI), which overcomes the distributional shift by constructing pessimistic estimates of the value functions for both players and outputs a policy pair by solving a correlated coarse equilibrium based on the two value functions. Furthermore, we establish a data-dependent upper bound on the suboptimality which recovers a sublinear rate without the assumption on uniform coverage of the dataset. We also prove an information-theoretical lower bound, which shows our upper bound is nearly minimax optimal, which suggests that the data-dependent term is intrinsic. Our theoretical results also highlight a notion of "relative uncertainty", which characterizes the necessary and sufficient condition for achieving sample efficiency in offline MGs. To the best of our knowledge, we provide the first nearly minimax optimal result for offline MGs with function approximation.

[*]Equal contribution [1]Center for Data Science, Peking University [2]The Hong Kong University of Science and Technology [3]Fudan University [4]Key Laboratory of Machine Perception, MOE, School of Artificial Intelligence, Peking University [5]Google Research [6]Northwestern University [7]Yale University. Correspondence to: Han Zhong <hanzhong@stu,pku.edu.cn>, Zhaoran Wang <zhaoranwang@gmail.com>, Zhuoran Yang <zhuoran.yang@yale.edu>.

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

1. Introduction

Reinforcement learning (RL) has recently achieved tremendous empirical success, including Go (Silver et al., 2016; 2017), Poker (Brown and Sandholm, 2019), robotic control (Kober et al., 2013), and Dota (Berner et al., 2019), many of which involve multiple agents. RL system with multiple agents acting in a common environment is referred to as multi-agent RL (MARL) where each agent aims to maximize its own long-term return by interacting with the environment and other agents (Zhang et al., 2021). Two key components of these successes are *function approximation* and *efficient simulators*. For modern RL applications with large state spaces, function approximations such as neural networks are used to approximate the value functions or the policies and contributes to the generalization across different state-action pairs. Meanwhile, an efficient simulator serves as the environment which allows the agent to collect millions to billions of trajectories for the training process.

However, for various scenarios, e.g., healthcare (Pan et al., 2017) and auto-driving (Wang et al., 2018) where either collecting data is costly and risky, or online exploration is not possible (Fu et al., 2020), it is far more challenging to apply (MA)RL methods in a trial-and-error fashion. To tackle these issues, offline RL aims to learn a good policy from a pre-collected dataset without further interacting with the environment. Recently, there has been impressive progress in the theoretical understanding about single-agent offline RL (Jin et al., 2020b; Rashidinejad et al., 2021; Zanette et al., 2021; Xie et al., 2021; Yin and Wang, 2021; Uehara and Sun, 2021), indicating that pessimism is critical for designing provably efficient offline algorithms. More importantly, these works demonstrate that the necessary and sufficient condition for achieving sample efficiency in offline MDP is the single policy (optimal policy) coverage. That is, it suffices for the offline dataset to have good coverage over the trajectories induced by the optimal policy.

In offline MARL for zero-sum Markov games, agents are not only facing the challenges of unknown environments, function approximation, and the distributional shift between the behavior policy and the optimal policy, but also challenged by the sophisticated minimax optimization for equilibrium solving. Due to these challenges, theoretical understand-

ings of offline MARL remains elusive. In particular, the following questions remain open:

(i) Can we design sample-efficient equilibrium learning algorithms in offline MARL?
(ii) What is the necessary and sufficient condition for achieving sample efficiency in offline MARL?

To this end, focusing on the two-player zero-sum and finite-horizon Markov Game (MG) with linear function approximation, we provide positive answers to the above two questions. Our contribution is threefold:

- For the two-player zero-sum MG with linear function approximation, we propose a computationally efficient algorithm, dubbed as pessimistic minimax value iteration (PMVI), which features the pessimism mechanism.
- We introduce a new notion of "relative uncertainty", which depends on the offline dataset and $(\pi^*, \nu) \cup (\pi, \nu^*)$, where (π^*, ν^*) is an NE and (π, ν) are arbitrary policies. Furthermore, we prove that the suboptimality of PMVI can be bounded by *relative uncertainty* up to multiplicative factors involving the dimension and horizon, which further implies that "low relative uncertainty" is the *sufficient* condition for NE finding in the offline linear MGs setting. Meanwhile, by constructing a counterexample, we prove that, unlike the single-agent MDP where the single policy (optimal policy) coverage is enough, it is impossible to learn an approximate NE by the dataset only with the single policy pair (NE) coverage property.
- We also investigate the necessary condition for NE finding in the offline linear MGs setting. We demonstrate that the *low relative uncertainty* is exactly the *necessary* condition by showing that the relative uncertainty is the information-theoretic lower bound. This lower bound also indicates that PMVI achieves minimax optimality up to multiplicative factors involving the dimension and horizon.

In summary, we propose the first *computationally efficient* algorithm for offline linear MGs which is *minimax optimal* up to multiplicative factors involving the dimension and horizon. More importantly, we figure out that *low relative uncertainty* is the *necessary and sufficient* condition for achieving sample efficiency in offline linear MGs setup.

1.1. Related Work

There is a rich literature on MG (Shapley, 1953) and RL. Due to space constraint, we focus on reviewing the theoretical works on two-player zero-sum MG and offline RL.

Two-player zero-sum Markov game. There has been an impressive progress for online two-player zero-sum MGs, including the tabular MG (Bai and Jin, 2020; Xie et al., 2020; Bai et al., 2020; Liu et al., 2021), and MGs with linear function approximation (Xie et al., 2020; Chen et al., 2021). Beyond these two settings, Jin et al. (2021) and Huang et al. (2021) consider the two-player zero-sum MG with general function approximation and the proposed algorithms can further solve MGs with kernel function approximation, MGs with rich observations, and kernel feature selection. For offline sampling oracle, Abe and Kaneko (2020) considers offline policy evaluation under the strong uniform concentration assumption.

Offline RL. The study of the offline RL (also known as batch RL), has a long history. In the single-agent setting, the prior works typically require a strong dataset coverage assumption (Precup, 2000; Antos et al., 2008; Levine et al., 2020), which is impractical in general, particularly for the modern RL problems with large state spaces. Recently, Jin et al. (2020b) takes a step towards identifying the minimal dataset assumption that empower provably efficient offline learning. In particular, it shows that pessimism principle allows efficient offline learning under a much weaker assumption which only requires a sufficient coverage over the optimal policy. After Jin et al. (2020b), a line of work (Rashidinejad et al., 2021; Yin and Wang, 2021; Uehara et al., 2021; Zanette et al., 2021; Xie et al., 2021; Uehara and Sun, 2021) leverages the principle of pessimism to design offline RL algorithms, both in the tabular case and in the case with general function approximation. These methods are not only more robust to the violation of dataset coverage assumption, but also provide non-trivial theoretical understandings of the offline learning, which are of independent interests. Despite the rich literature on single-agent offline RL, the extension to the MARL is still challenging.

To the best of our knowledge, the current work on sample-efficient equilibrium finding in offline MARL is only Zhong et al. (2021) and Cui and Du (2022). In particular, Zhong et al. (2021) studies the general-sum MGs with leader-follower structure and aims to find the Stackelberg-Nash equilibrium, but we focus on finding the NE in two-player zero-sum MGs with symmetric players. Our work is most closely related to the concurrent work Cui and Du (2022), which we discuss in detail below.

Comparison with Cui and Du (2022). Up to now, the concurrent work Cui and Du (2022) seems to provide the only analysis on tabular two-player zero-sum MG in the offline setting. We comment the similarities and differences between two works as follows.

In terms of algorithms, both PMVI (Algorithm 1) in this paper and algorithms proposed in Cui and Du (2022) are

pessimism-type algorithms and computationally efficient. Since tabular MG is a special case of linear MG, our algorithm can naturally be applied to the tabular setting and achieve sample efficiency under the same coverage assumption.

In terms of theoretical results, our work can be compared to Cui and Du (2022) in the following aspects. First, both this work and Cui and Du (2022) figure out the necessary and sufficient condition for achieving sample efficiency in (linear) MGs. Specifically, we introduce a new notion of *relative uncertainty* and prove that the *low relative uncertainty* is the necessary and sufficient condition for achieving sample efficiency in (linear) MGs. Cui and Du (2022) proposes a similar notion called *unilateral concentration* and obtains similar results. Second, by constructing slightly different hard instances, both this work and Cui and Du (2022) show that the single policy (NE) coverage assumption is not enough for NE identification in MGs. Third, this work and Cui and Du (2022) achieve near-optimal results in the linear setting and tabular setting, respectively. Finally, the information-theoretic lower bound in Cui and Du (2022) can be implied by that for single-agent MDP. In contrast, our information-theoretic lower bound is construction-based and is a non-trivial extension from single-agent MDP.

2. Preliminaries

In this section, we formally formulate our problem, and introduce preliminary concepts used in our paper.

2.1. Two-Player Zero-Sum Markov Game

We consider a two-player zero-sum, finite-horizon MG where one agent (referred to as the max-player) aims to maximize the total reward while the other agent (referred to as the min-player) aims to minimize it. The game is defined as a tuple $\mathcal{M}(H, \mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, r, \mathbb{P})$ where H is the number of steps in each episode, $\mathcal{S}$ is the state space, $\mathcal{A}_1, \mathcal{A}_2$ are the action spaces of the two players, respectively, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is the transition kernel where $\mathbb{P}_h(\cdot|s,a,b)$ is the distribution of the next state given the state-action pair (s,a,b) at step h, $r = \{r_h(s,a,b)\}_{h=1}^H$ is the reward function[1], where $r_h(s,a,b) \in [0,1]$ is the reward given the state-action pair (s,a,b) at step h. We assume that for each episode, the game starts with a fixed initial state $x \in \mathcal{S}$ and it can be straightforwardly generalized to the case where the initial state is sampled from some fixed but unknown distribution.

Policy and Value functions. Let $\Delta(\mathcal{X})$ be the probability simplex over the set $\mathcal{X}$. A Markov policy of the max-player is a sequence of functions $\pi = \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A}_1)\}$ where $\pi_h(s)$ is the distribution of actions taken by the max-player given the current state s at step h. Similarly, we can define the Markov policy of the min-player by $\nu = \{\nu_h : \mathcal{S} \to \Delta(\mathcal{A}_2)\}$. Given a policy pair (π, ν), the value function $V_h^{\pi,\nu} : \mathcal{S} \to \mathbb{R}$ and the Q-value function $Q_h^{\pi,\nu} : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}$ at step h are defined by

$$V_h^{\pi,\nu}(s_h) := \mathbb{E}_{\pi,\nu}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'}) \middle| s_h\right],$$
$$Q_h^{\pi,\nu}(s_h, a_h, b_h) := \mathbb{E}_{\pi,\nu}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'}) \middle| s_h, a_h, b_h\right],$$

where the expectation is taken over the randomness of the environment and the policy pair. We define the Bellman operator $\mathbb{B}_h$ for any function $V : \mathcal{S} \to \mathbb{R}$ as

$$\begin{aligned} &\mathbb{B}_h V(s,a,b) \\ =&\mathbb{E}[r_h(s,a,b) + V(s_{h+1}) \mid (s_h, a_h, b_h) = (s,a,b)]. \end{aligned} \tag{2.1}$$

It is not difficult to verify that the value function and Q-value function satisfy the following Bellman equation:

$$Q_h^{\pi,\nu}(s,a,b) = (\mathbb{B}_h V_{h+1}^{\pi,\nu})(s,a,b). \tag{2.2}$$

2.2. Linear Markov Game

We consider a family of MGs whose reward functions and transition kernels possess a linear structure.

Assumption 2.1 (Linear MGs (Xie et al., 2020))**.** For each $(s,a,b) \in \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2$, and $h \in [H]$, we have

$$\begin{aligned} r_h(x,a,b) &= \phi(x,a,b)^\top \theta_h, \\ \mathbb{P}_h(\cdot \mid x,a,b) &= \phi(x,a,b)^\top \mu_h(\cdot), \end{aligned} \tag{2.3}$$

where $\phi : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}^d$ is a known feature map, $\theta_h \in \mathbb{R}^d$ is an unknown vector, $\mu_h = (\mu_h^{(i)})_{i \in [d]}$ is a vector of d unknown signed measure over $\mathcal{S}$. We further assume that $||\phi(\cdot,\cdot,\cdot)|| \le 1$, $||\theta_h|| \le \sqrt{d}$, and $||\mu_h(\mathcal{S})|| \le \sqrt{d}$ for all $h \in [H]$ where $||\cdot||$ is the ℓ_2-norm of vector.

With this assumption, we have the following result.

Lemma 2.2 (Linearity of Value Function)**.** *Under Assumption 2.1, for any policy pair (π, ν) and any $(x,a,b,h) \in \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \times [H]$, we have*

$$Q^{\pi,\nu}(x,a,b) = \langle \phi(x,a,b), w_h^{\pi,\nu} \rangle,$$

where $w_h^{\pi,\nu} = \theta_h + \int_{\mathcal{S}} V_{h+1}^{\pi,\nu}(x') d\mu_h(x')$.

Proof. The result is implied by Bellman equation in (2.2) and the linearity of r_h and $\mathbb{P}_h$ in Assumption 2.1. □

[1] For ease of presentation, we consider deterministic reward. Our results immediately generalize to the stochastic reward function case.

2.3. Nash Equilibrium and Performance Metrics

If we fix some max-player's policy π, then the MG degenerates to an MDP for the min-player. By the theory of single-agent RL, we know that there exists a policy $\mathrm{br}(\pi)$, referred to as the best response policy of the min-player, satisfying $V_h^{\pi,\mathrm{br}(\pi)}(s) = \inf_\nu V_h^{\pi,\nu}(s)$ for all s and h. Similarly, we define the best response policy $\mathrm{br}(\nu)$ for the min-player's policy ν. To simplify the notation, we define

$$V_h^{\pi,*} = V_h^{\pi,\mathrm{br}(\pi)}, \text{ and } V_h^{*,\nu} \quad = V_h^{\mathrm{br}(\nu),\nu}.$$

It is known that there exists a Nash equilibrium (NE) policy (π^*, ν^*) such that π^* and ν^* are the best response policy to each other (Filar and Vrieze, 2012) and we denote the value of them as $V_h^* = V_h^{\pi^*,\nu^*}$. Although multiple NE policies may exist, for zero-sum MGs, the value function is unique.

The NE policy is further known to be the solution to the following minimax equation:

$$\sup_\pi \inf_\nu V_h^{\pi,\nu}(s) = V_h^{\pi^\star,\nu^\star}(s) = \inf_\nu \sup_\pi V_h^{\pi,\nu}(s), \forall (s,h). \tag{2.4}$$

We also have the following weak duality property for any policy pair (π, ν) in MG:

$$V_h^{\pi,*}(s) \le V_h^*(s) \le V_h^{*,\nu}(s), \forall (s,h). \tag{2.5}$$

Accordingly, we measure a policy pair (π, ν) by the duality gap:

$$\mathrm{SubOpt}((\pi,\nu), x) = V_1^{*,\nu}(x) - V_1^{\pi,*}(x). \tag{2.6}$$

The goal of learning is to find an ϵ-approximate NE $(\widehat{\pi}, \widehat{\nu})$ such that $\mathrm{SubOpt}((\widehat{\pi}, \widehat{\nu}), x) \le \epsilon$.

2.4. Offline Data Collecting Process

We introduce the notion of compliance of dataset.

Definition 2.3 (Compliance of Dataset)**.** Given an MG $\mathcal{M}$ and a dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, b_h^\tau)\}_{\tau,h=1}^{K,H}$, we say the dataset $\mathcal{D}$ is compliant with the MG $\mathcal{M}$ if

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}}\left(r_h^\tau = r, s_{h+1}^\tau = s | \{(s_h^i, a_h^i, b_h^i)\}_{i=1}^\tau, \{(r_h^i, s_{h+1}^i)\}_{i=1}^{\tau-1}\right)\\ &\quad = \mathbb{P}_h\left(r_h = r, s_{h+1} = s | s_h = s_h^\tau, a_h = a_h^\tau, b_h = b_h^\tau\right)\end{aligned} \tag{2.7}$$

for all $h \in [H], s \in \mathcal{S}$ where $\mathbb{P}$ in the right-hand side of (2.7) is taken with respect to the underlying MG $\mathcal{M}$.

We make the following assumption through this paper.

Assumption 2.4 (Date Collection)**.** The dataset $\mathcal{D}$ is compliant with the underlying MG $\mathcal{M}$.

Intuitively, the compliance ensures (i) $\mathcal{D}$ possesses the Markov property, and (ii) conditioned on $(s_h^\tau, a_h^\tau, b_h^\tau)$, (r_h^τ, s_{h+1}^τ) is generated by the reward function and the transition kernel of the underlying MG.

As discussed in Jin et al. (2020b), as a special case, this assumption holds if the dataset $\mathcal{D}$ is collected by a fixed behavior policy. More generally, the experimenter can sequentially improve her policy by any online MARL algorithm as the assumption allows (a_h^τ, b_h^τ) to be interdependent across the trajectories. In an extreme case, the actions can even be chosen in an adversarial manner.

2.5. Additional Notations

For any real number x and positive integer h, we define the regulation operation as $\Pi_h(x) = \min\{h, \max\{x, 0\}\}$. Given a semi-definite matrix Λ, the matrix norm for any vector v is denoted as $\|v\|_\Lambda = \sqrt{v^\top \Lambda v}$. The Frobenius norm of a matrix A is given by $||A||_F = \sqrt{\mathrm{tr}(AA^\top)}$. We denote $\lambda_{\min}(A)$ as the smallest eigenvalue of the matrix A. We also use the shorthand notations $\phi_h = \phi(s_h, a_h, b_h)$, $\phi_h^\tau = \phi(s_h^\tau, a_h^\tau, b_h^\tau)$, and $r_h^\tau = r_h(s_h^\tau, a_h^\tau, b_h^\tau)$.

3. Pessimistic Minimax Value Iteration

In this section, we introduce our algorithm, namely, Pessimistic Minimax Value Iteration (PMVI), whose peudocode is given in Algorithm 1.

Algorithm 1 Pessimistic Minimax Value Iteration

Input: Dataset $\mathcal{D} = \{x_h^\tau, a_h^\tau, b_h^\tau, r_h^\tau\}_{(\tau,h)\in[K]\times[H]}$.
Initialize $\underline{V}_{H+1}(\cdot) = \overline{V}_{H+1}(\cdot) = 0$.
for step $h = H, H-1, \cdots, 1$ **do**
  $\Lambda_h \leftarrow \sum_{\tau=1}^K \phi_h^\tau (\phi_h^\tau)^\top + I$.
  $\underline{w}_h \leftarrow \Lambda_h^{-1}(\sum_{\tau=1}^K \phi_h^\tau (r_h^\tau + \underline{V}_{h+1}(x_{h+1}^\tau)))$.
  $\overline{w}_h \leftarrow \Lambda_h^{-1}(\sum_{\tau=1}^K \phi_h^\tau (r_h^\tau + \overline{V}_{h+1}(x_{h+1}^\tau)))$.
  $\Gamma_h(\cdot,\cdot,\cdot) \leftarrow \beta \cdot \sqrt{\phi(\cdot,\cdot,\cdot)^\top (\Lambda_h)^{-1} \phi(\cdot,\cdot,\cdot)}$.
  $\underline{Q}_h(\cdot,\cdot,\cdot) \leftarrow \Pi_{H-h+1}\{\phi(\cdot,\cdot,\cdot)^\top \underline{w}_h - \Gamma_h(\cdot,\cdot,\cdot)\}$.
  $\overline{Q}_h(\cdot,\cdot,\cdot) \leftarrow \Pi_{H-h+1}\{\phi(\cdot,\cdot,\cdot)^\top \overline{w}_h + \Gamma_h(\cdot,\cdot,\cdot)\}$.
  Let $(\widehat{\pi}_h(\cdot \mid \cdot), \nu_h'(\cdot \mid \cdot))$ be the NE of the matrix game with payoff matrix $\underline{Q}_h(\cdot,\cdot,\cdot)$.
  Let $(\pi_h'(\cdot \mid \cdot), \widehat{\nu}_h(\cdot \mid \cdot))$ be the NE of the matrix game with payoff matrix $\overline{Q}_h(\cdot,\cdot,\cdot)$.
  $\underline{V}_h(\cdot) \leftarrow \mathbb{E}_{a\sim\widehat{\pi}_h(\cdot|\cdot), b\sim\nu_h'(\cdot|\cdot)} \underline{Q}_h(\cdot, a, b)$.
  $\overline{V}_h(\cdot) \leftarrow \mathbb{E}_{a\sim\pi_h'(\cdot|\cdot), b\sim\widehat{\nu}_h(\cdot|\cdot)} \overline{Q}_h(\cdot, a, b)$.
end for
Output: $(\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H, \widehat{\nu} = \{\widehat{\nu}_h\}_{h=1}^H)$.

At a high level, PMVI constructs pessimistic estimations of the value functions for both players and outputs a policy pair by solving a correlated coarse equilibrium based on these two estimated value functions.

Our learning process is done through backward induc-

tion with respect to the timestep h. We set $\underline{V}_{H+1}(\cdot) = \overline{V}_{H+1}(\cdot) = 0$, where $\underline{V}_{H+1}$ and $\overline{V}_{H+1}$ are estimated value functions for max-player and min-player, respectively. Suppose we have obtained the estimated value functions $(\underline{V}_{h+1}, \overline{V}_{h+1})$ at $(h+1)$-th step, together with the linearity of value functions (Lemma 2.2), we can use the regularized least-squares regression to obtain the linear coefficients $(\underline{w}_h, \overline{w}_h)$ for the estimated Q-functions:

$$\underline{w}_h \leftarrow \operatorname*{argmin}_w \sum_{\tau=1}^{K} [r_h^\tau + \underline{V}_{h+1}(x_{h+1}^\tau) - (\phi_h^\tau)^\top w]^2 + \|w\|_2^2,$$
$$\overline{w}_h \leftarrow \operatorname*{argmin}_w \sum_{\tau=1}^{K} [r_h^\tau + \overline{V}_{h+1}(x_{h+1}^\tau) - (\phi_h^\tau)^\top w]^2 + \|w\|_2^2,$$

where ϕ_h^τ is the shorthand of $\phi(s_h^\tau, a_h^\tau, b_h^\tau)$. Solving this problem gives the closed-form solutions:

$$\begin{aligned} \underline{w}_h &\leftarrow \Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi_h^\tau (r_h^\tau + \underline{V}_{h+1}(x_{h+1}^\tau))\Big), \\ \overline{w}_h &\leftarrow \Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi_h^\tau (r_h^\tau + \overline{V}_{h+1}(x_{h+1}^\tau))\Big), \\ \text{where } \Lambda_h &\leftarrow \sum_{\tau=1}^{K} \phi_h^\tau (\phi_h^\tau)^\top + I. \end{aligned} \tag{3.1}$$

Unlike the online setting where optimistic estimations are essential for encouraging exploration (Jin et al., 2020a; Xie et al., 2020), we need to adopt more robust estimation due to the distributional shift in the offline setting. Inspired by recent work (Jin et al., 2020b; Rashidinejad et al., 2021; Yin and Wang, 2021; Uehara and Sun, 2021; Zanette et al., 2021), which shows that *pessimism* plays a key role in the offline setting, we also use the pessimistic estimations for both players. In detail, we estimate Q-functions by subtracting/adding a bonus term:

$$\begin{aligned} \underline{Q}_h(\cdot,\cdot,\cdot) &\leftarrow \Pi_{H-h+1}\{\phi(\cdot,\cdot,\cdot)^\top \underline{w}_h - \Gamma_h(\cdot,\cdot,\cdot)\}, \\ \overline{Q}_h(\cdot,\cdot,\cdot) &\leftarrow \Pi_{H-h+1}\{\phi(\cdot,\cdot,\cdot)^\top \overline{w}_h + \Gamma_h(\cdot,\cdot,\cdot)\}. \end{aligned} \tag{3.2}$$

Here Γ_h is the bonus function, which takes the form $\beta\sqrt{\phi^\top \Lambda_h^{-1}\phi}$, where β is a parameter which will be specified later. Such a bonus function is common in linear bandits (Lattimore and Szepesvári, 2020) and linear MDPs (Jin et al., 2020a). We remark that $\underline{Q}_h$ and $\overline{Q}_h$ are pessimistic estimations for the max-player and the min-player, respectively. Then, we solve the matrix games with payoffs $\underline{Q}_h$ and $\overline{Q}_h$:

$$\begin{aligned} (\widehat{\pi}_h(\cdot \mid \cdot), \nu_h'(\cdot \mid \cdot)) &\leftarrow \mathrm{NE}(\underline{Q}_h(\cdot,\cdot,\cdot)), \\ (\pi_h'(\cdot \mid \cdot), \widehat{\nu}_h(\cdot \mid \cdot)) &\leftarrow \mathrm{NE}(\overline{Q}_h(\cdot,\cdot,\cdot)). \end{aligned}$$

The estimated value functions $\underline{V}_h(\cdot)$ and $\overline{V}_h(\cdot)$ are defined by $\mathbb{E}_{a\sim\widehat{\pi}_h(\cdot|\cdot), b\sim\nu_h'(\cdot|\cdot)}\underline{Q}_h(\cdot,a,b)$ and $\mathbb{E}_{a\sim\pi_h'(\cdot|\cdot), b\sim\widehat{\nu}_h(\cdot|\cdot)}\overline{Q}_h(\cdot,a,b)$, respectively. After H steps, PMVI outputs the policy pair $(\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H, \widehat{\nu} = \{\widehat{\nu}_h\}_{h=1}^H)$.

Remark 3.1 (Computational efficiency). We remark that our algorithm is computationally efficient because both the regression (3.1) and finding the NE of a zero-sum matrix game (using linear programming) can be efficiently implemented. Moreover, we remark that we do not need to compute $\overline{Q}_h(x,\cdot,\cdot), \underline{Q}_h(x,\cdot,\cdot), \widehat{\pi}_h(\cdot|x), \nu_h'(\cdot|x), \pi_h'(\cdot|x), \widehat{\nu}_h'(\cdot|x)$ for all $x \in \mathcal{S}$. Instead, we only do so for the states we encounter.

Remark 3.2. We remark that the linearity of the reward functions and the transition kernel is *strictly* stronger than the linearity of value-function. In the online setting, the recent works (Jin et al., 2021; Huang et al., 2021) show that the linearity of the value function empowers *statistically* efficient learning. However, we consider this stronger assumption because it is likely that it is essential for *computational* efficiency due to the lack of computation tractability with general function approximation and the hardness result in Du et al. (2019) which only assumes near-linearity of value functions of MDPs (special case of MGs).

In the following theorem, we provide the theoretical guarantees for PMVI (Algorithm 1). Recall that we use the shorthand $\phi_h = \phi(s_h, a_h, b_h)$.

Theorem 3.3. *Suppose Assumptions 2.1 and 2.4 hold. Set $\beta = cdH\sqrt{\zeta}$ in Algorithm 1, where c is a sufficient large constant and $\zeta = \log(2dKH/p)$. Then for sufficient large K, it holds with probability $1-p$ that*

$$\begin{aligned} &\mathrm{SubOpt}\left((\widehat{\pi}, \widehat{\nu}), x\right) \\ &\le 2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu'}\left[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}\,\middle|\, s_1 = x\right] \\ &\quad + 2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}\left[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}\,\middle|\, s_1 = x\right]. \end{aligned}$$

Proof. See Appendix A for a detailed proof. □

Theorem 3.3 states that the suboptimality of PMVI is upper bounded by the product of 2β and a data-dependent term, where β comes from the the covering number of function classes and the date-dependent term will be explained in the following section.

4. Sufficiency: Low Relative Uncertainty

In this section, we interpret Theorem 3.3 by characterizing the sufficient condition for achieving sample efficiency.

4.1. Relative Uncertainty

We first introduce the following important notion of "relative uncertainty".

Definition 4.1 (Relative Uncertainty)**.** Given an MG $\mathcal{M}$ and a dataset $\mathcal{D}$ that is compliant with $\mathcal{M}$, for an NE policy pair (π^*, ν^*), the relative uncertainty of (π^*, ν^*) with respect to $\mathcal{D}$ is defined as

$$\begin{aligned}&\mathrm{RU}(\mathcal{D}, \pi^*, \nu^*, x)\\ &= \max\Big\{\sup_{\nu}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\nu}\Big[\sqrt{\phi_h^\top\Lambda_h^{-1}\phi_h}\,\Big|\, s_1 = x\Big],\\ &\qquad \sup_{\pi}\sum_{h=1}^{H}\mathbb{E}_{\pi,\nu^*}\Big[\sqrt{\phi_h^\top\Lambda_h^{-1}\phi_h}\,\Big|\, s_1 = x\Big]\Big\},\end{aligned}$$

where x is the initial state and expectation $\mathbb{E}_{\pi^*,\nu}$ and $\mathbb{E}_{\pi,\nu^*}$ are taken respect to randomness of the trajectory induced by (π^*, ν) and (π, ν^*) in the underlying MG given the fixed matrix $\Lambda_h = \sum_{\tau=1}^{K}\phi_h^\tau(\phi_h^\tau)^\top + I$, respectively.

We also define the relative uncertainty with respect to the dataset $\mathcal{D}$ as

$$\mathrm{RU}(\mathcal{D}, x) = \inf_{(\pi^*,\nu^*)\text{ is NE}} \mathrm{RU}(\mathcal{D}, \pi^*, \nu^*, x). \tag{4.1}$$

Therefore, we can reformulate Theorem 3.3 as:

$$\mathrm{SubOpt}((\widehat{\pi}, \widehat{\nu}), x) \le 4\beta \cdot \mathrm{RU}(\mathcal{D}, x). \tag{4.2}$$

Hence, we obtain that "low relative uncertainty" allows PMVI to find an approximate NE policy pair sample efficiently, which further implies that "low relative uncertainty" is the sufficient condition for achieving sample efficiency in offline linear MGs.

Before we provide a detailed discussion of this notion with intuitions and examples, we first contrast our result with the single policy (optimal policy) coverage identified in the single-agent setting (Jin et al., 2020b; Xie et al., 2021; Rashidinejad et al., 2021).

4.2. Single Policy (NE) Coverage is Insufficient

As demonstrated in Jin et al. (2020b); Xie et al. (2021); Rashidinejad et al. (2021), a sufficient coverage over the optimal policy is sufficient for the offline learning of MDPs. As a straightforward extension, it is natural to ask whether a sufficient coverage over the NE policy pair (π^*, ν^*) is sufficient and therefore minimal. However, the situation is more complicated in the MG case and we have the following impossibility result.

Proposition 4.2. *Coverage of the NE policy pair (π^*, ν^*) is not sufficient for learning an approximate NE policy pair.*

Proof. We prove the result by constructing two hard instances and a dataset $\mathcal{D}$ such that no algorithm can achieve small suboptimality for two instances simultaneously. We consider two simplified linear MGs $\mathcal{M}_1$ and $\mathcal{M}_2$ with state space $\mathcal{S} = \{X\}$, action sets $\mathcal{A}_1 = \{a_i : i \in [3]\}$, $\mathcal{A}_2 = \{b_i : i \in [3]\}$, and payoff matrices:

$$R_1 = \begin{pmatrix} 0.5 & -1 & 0\\ 1 & 0 & 1\\ 0 & -1 & 0\end{pmatrix},\quad R_2 = \begin{pmatrix} 0 & 0 & -1\\ 1 & 0 & -1\\ 1 & 1 & 0\end{pmatrix}. \tag{4.3}$$

We consider the dataset $\mathcal{D} = \{(a_2, b_2, r = 0), (a_3, b_3, r = 0)\}$ where the choices of action are predetermined and the rewards are sampled from the underlying game, which implies that $\mathcal{D}$ is compliant with the underlying game. However, we can never distinguish these two games as they are both consistent with $\mathcal{D}$. Suppose that the output policies are $\widehat{\pi}(a_i) = p_i, \widehat{\nu}(b_j) = q_j$ with $i, j \in [3]$, we can easily find that

$$\begin{aligned}\mathrm{SubOpt}_{\mathcal{M}_1}((\widehat{\pi}, \widehat{\nu}), x) &= 2 - p_2 - q_2,\\ \mathrm{SubOpt}_{\mathcal{M}_2}((\widehat{\pi}, \widehat{\nu}), x) &= p_1 + q_1 + p_2 + q_2,\end{aligned}$$

where the subscript $\mathcal{M}_i$ means that the underlying MG is $\mathcal{M}_i$. Therefore, we have

$$\mathrm{SubOpt}_{\mathcal{M}_1}((\widehat{\pi}, \widehat{\nu}), x) + \mathrm{SubOpt}_{\mathcal{M}_2}((\widehat{\pi}, \widehat{\nu}), x) \ge 2,$$

which implies that either $\mathrm{SubOpt}_{\mathcal{M}_1}((\widehat{\pi}, \widehat{\nu}), x)$ or $\mathrm{SubOpt}_{\mathcal{M}_2}((\widehat{\pi}, \widehat{\nu}), x)$ is larger than 1. □

We remark that the instances constructed in the proof also intuitively illustrate the sufficiency of the "low relative uncertainty". Suppose that the underlying MG is $\mathcal{M}_1$ defined in (4.3) and the dataset $\mathcal{D}$ now contains the information about the set of action pairs:

$$G = \{(a_1, b_2), (a_2, b_2), (a_2, b_1), (a_2, b_3), (a_3, b_2)\}. \tag{4.4}$$

Then, the learning agent has the following estimation

$$\widehat{R} = \begin{pmatrix} * & -1 & *\\ 1 & 0 & 1\\ * & -1 & *\end{pmatrix}, \tag{4.5}$$

where $*$ can be arbitrary. In particular, the collected information is sufficient to verify that (a_2, b_2) are best response to each other and therefore the NE policy pair.

More generally, for the NE that is possible a mixed strategy, if we have sufficient information about $\{(\pi^*, \nu) : \nu \text{ is arbitrary}\}$, we can verify that ν^* is the best response of π^*. Similarly, the information about $\{(\pi, \nu^*) :$

π is arbitrary$\}$ allows us to ensure that π^* is the best response policy to ν^*. Therefore, intuitively, a sufficient coverage over these policy pairs empowers efficient offline learning of the NE.

4.3. Interpretation of Theorem 3.3

To illustrate our theory more, we make several comments below.

Data-Dependent Performance Upper Bound. The upper bound in Theorem 3.3 is in a data-dependent manner, which is also a key idea employed by many previous works. This allows to drop the strong uniform coverage assumption, which usually fails to hold in practice. Specifically, the suboptimality guarantee only relies on the the compliance assumption and depends on the dataset $\mathcal{D}$ through the relative uncertainty $\mathrm{RU}(\mathcal{D}, x)$.

To better illustrate the role of the relative uncertainty, we consider the linear MG $\mathcal{M}_1$ constructed in (4.3). We define n_{ij} as the times that (a_i, b_j) is taken in $\mathcal{D}$. Then, we have

$$\sup_{\nu} \mathbb{E}_{\pi^*,\nu}\left[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}\middle| s_1 = x\right] = (1+\min_j n_{2,j})^{-1/2},$$
$$\sup_{\pi} \mathbb{E}_{\pi,\nu^*}\left[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}\middle| s_1 = x\right] = (1+\min_i n_{i,2})^{-1/2},$$

which implies that

$$\mathrm{RU}(\mathcal{D}, x) = \mathrm{RU}(\mathcal{D}, \pi^*, \nu^*, x) = (1+n^*)^{-1/2}, \quad (4.6)$$

where $n^* = \min_{i,j\in[3]}\{n_{2,j}, n_{i,2}\}$. Hence, $\mathrm{RU}(\mathcal{D}, x)$ measures how well the dataset $\mathcal{D}$ covers the action pairs induced by (π^*, ν) and (π, ν^*), where π and ν are arbitrary. In particular, combining (4.2) and (4.6), we obtain that

$$\mathrm{SubOpt}((\widehat{\pi}, \widehat{\nu}), x) \le 4\beta \cdot (1+n^*)^{-1/2},$$

where we take β as stated in the theorem. This implies that the suboptimality of Algorithm 1 is small if the action pair set is covered well by $\mathcal{D}$, which corresponds to a large n^*. More generally, we have the following corollary:

Corollary 4.3 (Sufficient Coverage of Relative Information)**.** *Under Assumptions 2.1 and 2.4, we assume the existence of a constant c_1 such that*

$$\Lambda_h \geqslant I + c_1 \cdot K \cdot \max\Big\{\sup_{\nu} \mathbb{E}_{\pi^*,\nu}\left[\phi_h\phi_h^\top \middle| s_1 = x\right], \sup_{\pi} \mathbb{E}_{\pi,\nu^*}\left[\phi_h\phi_h^\top \middle| s_1 = x\right]\Big\}, \quad (4.7)$$

with probability at least $1-p/2$. Set $\beta = cdH\sqrt{\zeta}$ in Algorithm 1 where c is a sufficient large constant and $\zeta = \log(4dHK/p)$. Then for sufficient large K, it holds with probability $1-p$ that

$$\mathrm{SubOpt}((\widehat{\pi}, \widehat{\nu}), x) \leqslant c' d^{3/2} H^2 K^{-1/2}\sqrt{\zeta},$$

where c' is a constant that only relies on c and c_1.

Proof. See Appendix B for detailed proof. □

Oracle Property. Notably, in the above example, the action pair that lies off the set G in (4.4) will not affect $\mathrm{RU}(\mathcal{D}, x)$. Such a property is referred as the oracle property in the literature (Donoho and Johnstone, 1994; Zou, 2006; Fan and Li, 2001). Specifically, since $\mathrm{RU}(\mathcal{D}, x)$ takes expectation under the set of policy pairs:

$$P = \{(\pi^*, \nu) : \nu \text{ is arbitrary}\} \bigcup \{(\pi, \nu^*) : \pi \text{ is arbitrary}\},$$

the suboptimality automatically "adapts" to the trajectory induced by this set even though it is unknown in prior. This property is important especially when the dataset $\mathcal{D}$ contains a large amount of irrelative information as the irrelative information possibly misleads other learning agents. For instance, suppose that we collect $\mathcal{D}$ through a naive policy pair where both the max-player and the min-player pick their actions randomly. Therefore, all action pairs are sampled approximately uniformly. We assume that they are equally sampled for $K/9$ times for simplicity. In this case, since $n^* = K/9$, the suboptimality of Algorithm 1 still decays at a rate of $K^{-1/2}$. In particular, one important observation is that the output policy pair $(\widehat{\pi}, \widehat{\nu})$ can outperform the naive policy used to collect the dataset $\mathcal{D}$.

Well-Explored Dataset. As in existing literature (e.g., (Duan et al., 2020)), we also consider the case where the data collecting process explores the state-action space well.

Corollary 4.4 (Well-Explored Dataset)**.** *Supposed the dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, b_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ is induced by a fixed behavior policy pair $(\overline{\pi}, \overline{\nu})$ in the underlying MG. We also assume the existence of a constant $\underline{c} > 0$ such that*

$$\lambda_{min}(\Sigma_h) \geqslant \underline{c} \quad \textit{where} \quad \Sigma_h = \mathbb{E}_{\overline{\pi},\overline{\nu}}[\phi_h\phi_h^\top], \quad \forall h \in [H].$$

Set $\beta = cdH\sqrt{\zeta}$ in Algorithm 1 where c is a sufficient large constant and $\zeta = \log(4dHK/p)$. Then for sufficient large K, it holds with probability $1-p$ that

$$\mathrm{SubOpt}((\widehat{\pi}, \widehat{\nu}), x) \leqslant c' d H^2 K^{-1/2}\sqrt{\zeta},$$

where c' is a constant that only relies on c and $\underline{c}$.

Proof. See Appendix C for a detailed proof. □

5. Necessity: Low Relative Uncertainty

In this section, we show that the low relative uncertainty is also the necessary condition by establishing an information-theoretic lower bound.

We have considered two sets of policy pairs, corresponding to two levels of coverage assumptions on the dataset:

$$P_1 = \{(\pi^*, \nu^*) \text{ is an NE}\};$$
$$P_2 = \{(\pi^*, \nu), (\pi, \nu^*) : \pi, \nu \text{ are arbitrary}\}.$$

Clearly, we have $P_1 \subset P_2$. From the discussion in Section 4, we know that a good coverage of P_1 is insufficient, while a good coverage over P_2 is sufficient for efficient offline learning. It remains to ask whether there is a coverage assumption weaker than P_2 but stronger than P_1 that empowers efficient offline learning in our setting. We give the negative answer by providing an information-theoretic lower bound in the following theorem.

Theorem 5.1. *For any algorithm* $\mathrm{Algo}(\cdot)$ *that outputs a Markov policy pair based on* $\mathcal{D}$*, there exists a linear game* $\mathcal{M}$ *and a dataset* $\mathcal{D}$ *that is compliant with the underlying MG* $\mathcal{M}$*, such that when* K *is large enough, it holds that*

$$\mathbb{E}_{\mathcal{D}}\left[\frac{\mathrm{SubOpt}\left(\mathrm{Algo}(\mathcal{D}); x_0\right)}{\mathrm{RU}(\mathcal{D}, x_0)}\right] \geqslant C', \tag{5.1}$$

where C' *is an absolute constant and* x_0 *is the initial state. The expectation is taken with respect to* $\mathbb{P}_{\mathcal{D}}$ *where* $\mathrm{Algo}(\mathcal{D})$ *is a policy pair constructed based on the dataset* $\mathcal{D}$*.*

Proof. See Appendix D for a detailed proof. □

Notably, the lower bound in Theorem 5.1 matches the suboptimality upper bound in Theorem 3.3 up to β and absolute constant factors and therefore establishes the near-optimality of Algorithm 1. Meanwhile, Theorem 5.1 states that the relative uncertainty $\mathrm{RU}(\mathcal{D}, x_0)$ correctly captures the hardness of offline MG under the linear function approximation setting, that is, low relative uncertainty is the necessary condition for achieving sample efficiency.

6. Conclusion

In this paper, we make the first attempt to study the two-player zero-sum linear MGs in the offline setting. For such an equilibrium finding problem, we propose a pessimism-based algorithm PMVI, which is the first RL algorithm that can achieve both computational efficiency and nearly minimax optimality. Meanwhile, we introduce a new notion of relative uncertainty and prove that low relative uncertainty is the necessary and sufficient condition for achieving sample efficiency in offline linear MGs. We believe our work opens up many promising directions for future work, such as how to perform sample-efficient equilibrium learning in the offline zero-sum MGs with general function approximations (Jin et al., 2021; Huang et al., 2021).

Acknowledgement

The authors would like to thank Qiaomin Xie for helpful discussions. Liwei Wang was supported by National Key R&D Program of China (2018YFB1402600), Exploratory Research Project of Zhejiang Lab (No. 2022RC0AN02), BJNSF (L172037), Project 2020BD006 supported by PKUBaidu Fund, the major key project of PCL (PCL2021A12). Wei Xiong and Tong Zhang acknowledge the funding supported by GRF 16201320 and the Hong Kong Ph.D. Fellowship.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

Kenshi Abe and Yusuke Kaneko. Off-policy exploitability-evaluation in two-player zero-sum markov games. *arXiv preprint arXiv:2007.02141*, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

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

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

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Dkebiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

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

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021.

Qiwen Cui and Simon S Du. When is offline two-player zero-sum markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022.

David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

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

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020a.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020b.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Yunpeng Pan, Ching-An Cheng, Kamil Saigol, Keuntaek Lee, Xinyan Yan, Evangelos Theodorou, and Byron Boots. Agile autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*, 2017.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587): 484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Joel A. Tropp. An introduction to matrix concentration inequalities, 2015.

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

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

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

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2447–2456, 2018.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pages 3674–3682. PMLR, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

Bin Yu, Fano Assouad, and Lucien Le Cam. Festschrift for lucien le cam, 1997.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

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

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476): 1418–1429, 2006.

A. Proof of Theorem 3.3

Proof of Theorem 3.3. First, we define the Bellman error as

$$\underline{\iota}_h(x,a,b) = \mathbb{B}_h\underline{V}_{h+1}(x,a,b) - \underline{Q}_h(x,a,b),$$
$$\overline{\iota}_h(x,a,b) = \mathbb{B}_h\overline{V}_{h+1}(x,a,b) - \overline{Q}_h(x,a,b).$$

Our proof relies on the following lemma.

Lemma A.1. *Let $\mathcal{E}$ denote the event that*

$$0 \le -\overline{\iota}_h(s,a,b) \le 2\Gamma_h(s,a,b),$$
$$0 \le \underline{\iota}_h(s,a,b) \le 2\Gamma_h(s,a,b).$$

for all $h \in [H]$ and $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$. Then we have $\Pr(\mathcal{E}) \ge 1-p$.

Proof. See Appendix A.1 for a detailed proof. □

Under this event, we also have the following lemma to ensure that our estimated value functions are optimistic.

Lemma A.2. *Under the event $\mathcal{E}$, we have*

$$\underline{V}_h(x) \le V_h^{\widehat{\pi},*}(x), \qquad V_h^{*,\widehat{\nu}}(x) \le \overline{V}_h(x).$$

Proof. See Appendix A.2 for a detailed proof. □

Back to our proof, we decompose the suboptimality gap as

$$\mathrm{SubOpt}\left((\widehat{\pi},\widehat{\nu}),x\right) = V_1^{*,\widehat{\nu}}(x) - V_1^{\widehat{\pi},*}(x) = \underbrace{V_1^{*,\widehat{\nu}}(x) - V_1^{*}(x)}_{\text{(i)}} + \underbrace{V_1^{*}(x) - V_1^{\widehat{\pi},*}(x)}_{\text{(ii)}}. \tag{A.1}$$

For term (i), by Lemma A.2, we have

$$\text{(i)} \le \overline{V}_1(x) - V_1^{*}(x) \le \overline{V}_1(x) - V_1^{\pi',\nu^*}(x), \tag{A.2}$$

where the last inequality follows from the fact that (π^*,ν^*) is the NE. Then we can use the following lemma to decompose the term $\overline{V}_1(x) - V_1^{\pi',\nu^*}(x)$.

Lemma A.3 (Value Difference Lemma)**.** *Given an MG $(\mathcal{S},\mathcal{A},\mathcal{B},r,H)$. Let $\widehat{\pi}\otimes\widehat{\nu} = \{\widehat{\pi}_h\otimes\widehat{\nu}_h : \mathcal{S}\to\Delta(\mathcal{A}_1)\times\Delta(\mathcal{A}_2)\}_{h\in[H]}$ be a product policy, (π,ν) be a policy pair, and $\{\widehat{Q}_h\}_{h=1}^H$ be any estimated Q-functions. For any $h\in[H]$, we define the estimated value function $\widehat{V}_h : \mathcal{S}\to\mathbb{R}$ by setting $\widehat{V}_h(x) = \langle\widehat{Q}_h(x,\cdot,\cdot),\widehat{\pi}_h(\cdot|x)\otimes\widehat{\nu}_h(\cdot|x)\rangle$ for all $x\in\mathcal{S}$. For all $x\in\mathcal{S}$,*

$$\widehat{V}_1(x) - V_1^{\pi,\nu}(x) = \sum_{h=1}^H \mathbb{E}_{\pi,\nu}\Big[\langle\widehat{Q}_h(s_h,\cdot,\cdot),\widehat{\pi}_h(\cdot|s_h)\otimes\widehat{\nu}_h(\cdot|s_h) - \pi_h(\cdot|s_h)\otimes\nu_h(\cdot|s_h)\rangle|s_1 = x\Big]$$
$$+\sum_{h=1}^H \mathbb{E}_{\pi,\nu}\left[\widehat{Q}_h(s_h,a_h,b_h) - \mathbb{B}_h\widehat{V}_{h+1}(s_h,a_h,b_h)|s_1=x\right].$$

Proof. See Section B.1 in Cai et al. (2020) for a detailed proof. □

By Lemma A.3, we obtain

$$\overline{V}_1(x) - V_1^{\pi',\nu^*}(x) = \sum_{h=1}^H \mathbb{E}_{\pi',\nu^*}\Big[\langle\overline{Q}_h(s_h,\cdot,\cdot),\pi_h'(\cdot|x)\otimes\widehat{\nu}_h(\cdot|x) - \pi_h'(\cdot|s_h)\otimes\nu_h^*(\cdot|s_h)\rangle|s_1=x\Big]$$
$$-\sum_{h=1}^H \mathbb{E}_{\pi',\nu^*}[\overline{\iota}_h(s_h,a_h,b_h)|s_1=x]. \tag{A.3}$$

The first term can be bounded by the following lemma.

Lemma A.4. *It holds that*

$$\sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}\Big[\langle \overline{Q}_h(s_h,\cdot,\cdot), \pi'_h(\cdot|s_h)\otimes\widehat{\nu}_h(\cdot|s_h) - \pi'_h(\cdot|s_h)\otimes\nu^*_h(\cdot|s_h)\rangle | s_1 = x\Big] \le 0.$$

Proof. See Appendix A.3 for a detailed proof. □

Applying Lemma A.1 to the second term of (A.3) gives

$$-\sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}[\underline{\iota}_h(s_h,a_h,b_h)|s_1=x] \le 2\sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}[\Gamma_h(s_h,a_h,b_h)|s_1=x].$$

Putting the above inequalities together we obtain

$$\text{(i)} \le 2\sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*}[\Gamma_h(s_h,a_h,b_h)|s_1=x] \tag{A.4}$$

Similarly, we can obtain

$$\text{(ii)} \le 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu'}[\Gamma_h(s_h,a_h,b_h)|s_1=x]. \tag{A.5}$$

Plugging (A.4) and (A.5) into (A.1), we conclude the proof of Theorem 3.3. □

A.1. Proof of Lemma A.1

Proof of Lemma A.1. Throughout this proof, we use the shorthands

$$\phi_h^\tau = \phi(s_h^\tau, a_h^\tau, b_h^\tau), \qquad \phi_h = \phi(s_h,a_h,b_h), \qquad \phi = \phi(s,a,b).$$

For the simplicity of notation, we also let

$$\epsilon_h^\tau(V) = r_h^\tau + V(s_{h+1}^\tau) - \mathbb{B}_h V(s_h^\tau, a_h^\tau, b_h^\tau) \tag{A.6}$$

By the linear MDP assumption, we have $(\mathbb{B}_h\overline{V}_{h+1})(s,a,b) = \phi(s,a,b)^\top w_h$, where

$$w_h = \theta_h + \int_{x\in\mathcal{S}} \overline{V}_{h+1}(x)\mu_h(x)\mathrm{d}x.$$

Then we have

$$
\begin{aligned}
&\left|\phi^\top \overline{w}_h - (\mathbb{B}_h \overline{V}_{h+1})(s,a,b)\right| \\
&\quad= \left|\phi^\top \left(\Lambda_h^{-1} \sum_{\tau=1}^{K} \left(r_h^\tau + \overline{V}_{h+1}(s_{h+1}^\tau)\right) \phi_h^\tau\right) - (\mathbb{B}_h \overline{V}_{h+1})(s,a,b)\right| \\
&\quad= \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} \epsilon_h^\tau(\overline{V}_{h+1})\phi_h^\tau + \phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} (\mathbb{B}_h \overline{V}_{h+1})(s_h^\tau, a_h^\tau, b_h^\tau)\phi_h^\tau - \phi^\top w_h\right| \\
&\quad= \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} \epsilon_h^\tau(\overline{V}_{h+1})\phi_h^\tau + \phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} \phi_h^\tau (\phi_h^\tau)^\top w_h - \phi^\top w_h\right| \\
&\quad= \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} \epsilon_h^\tau(\overline{V}_{h+1})\phi_h^\tau + \phi^\top \Lambda_h^{-1} (\Lambda_h - I) w_h - \phi^\top w_h\right| \\
&\quad= \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} \epsilon_h^\tau(\overline{V}_{h+1})\phi_h^\tau - \phi^\top \Lambda_h^{-1} w_h\right| \\
&\quad\leqslant \underbrace{\left|\phi^\top \Lambda_h^{-1} w_h\right|}_{\text{(i)}} + \underbrace{\left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} \epsilon_h^\tau(\overline{V}_{h+1})\phi_h^\tau\right|}_{\text{(ii)}}.
\end{aligned} \tag{A.7}
$$

Now we estimate term (i)

$$
|\text{(i)}| \leqslant \|\phi\|_{\Lambda_h^{-1}} \|w_h\|_{\Lambda_h^{-1}} \leqslant \|w_h\| \|\phi\|_{\Lambda_h^{-1}} \leqslant H\sqrt{d} \|\phi\|_{\Lambda_h^{-1}}, \tag{A.8}
$$

where the second inequality follows from $\left\|\Lambda_h^{-1}\right\|_{\text{op}} \leqslant 1$ and the third inequality follows from Lemma E.1. Here $\|\cdot\|_{\text{op}}$ denotes the operator norm of a matrix.

Supposed that $\|V - \overline{V}_{h+1}\|_\infty \leqslant \epsilon$, by the definition of $\epsilon_h^\tau(V)$ in (A.6), we have

$$
\begin{aligned}
&\left|\epsilon_h^\tau(\overline{V}_{h+1}) - \epsilon_h^\tau(V)\right| \\
&\quad= \left|r_h^\tau + \overline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{B}_h \overline{V}_{h+1}(s_h^\tau, a_h^\tau, b_h^\tau) - r_h^\tau - V(s_{h+1}^\tau) + \mathbb{B}_h V(s_h^\tau, a_h^\tau, b_h^\tau)\right| \\
&\quad\leqslant \left|\overline{V}_{h+1}(s_{h+1}^\tau) - V(s_{h+1}^\tau)\right| + \left|\mathbb{B}_h \overline{V}_{h+1}(s_h^\tau, a_h^\tau, b_h^\tau) - \mathbb{B}_h V(s_h^\tau, a_h^\tau, b_h^\tau)\right| \leqslant 2\epsilon.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
\left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^{K} \left(\epsilon_h^\tau(\overline{V}_{h+1}) - \epsilon_h^\tau(V)\right) \phi_h^\tau\right| &\leqslant \sum_{\tau=1}^{K} \left|\phi^\top \Lambda_h^{-1} \left(\epsilon_h^\tau(\overline{V}_{h+1}) - \epsilon_h^\tau(V)\right) \phi_h^\tau\right| \\
&\leqslant \sum_{\tau=1}^{K} \left|\epsilon_h^\tau(\overline{V}_{h+1}) - \epsilon_h^\tau(V)\right| \|\phi\|_{\Lambda_h^{-1}} \|\phi_h^\tau\|_{\Lambda_h^{-1}} \\
&\leqslant \sum_{\tau=1}^{K} \left|\epsilon_h^\tau(\overline{V}_{h+1}) - \epsilon_h^\tau(V)\right| \|\phi\|_{\Lambda_h^{-1}} \|\phi_h^\tau\| \leqslant 2\epsilon K \|\phi\|_{\Lambda_h^{-1}},
\end{aligned}
$$

where the last inequality holds since $\|\phi\| \leqslant 1$. We define two function classes as

$$
\begin{aligned}
\underline{\mathcal{Q}}_h &= \Pi_{H-h+1}\left\{\phi(\cdot,\cdot,\cdot)^\top w - \beta\sqrt{\phi^\top \Lambda^{-1} \phi}\right\}, \\
\overline{\mathcal{Q}}_h &= \Pi_{H-h+1}\left\{\phi(\cdot,\cdot,\cdot)^\top w + \beta\sqrt{\phi^\top \Lambda^{-1} \phi}\right\},
\end{aligned} \tag{A.9}
$$

where the parameters (w, Λ) satisfy $\|w\| \leq H\sqrt{dK}$ and $\lambda_{\min}(\Lambda) \geq 1$. Let $\underline{\mathcal{Q}}_{h,\epsilon}$ and $\overline{\mathcal{Q}}_{h,\epsilon}$ be the ϵ-nets of $\underline{\mathcal{Q}}_h$ and $\overline{\mathcal{Q}}_h$, respectively. Choose the pair $(\underline{Q}'_{h+1}, \overline{Q}'_{h+1}) \in \underline{\mathcal{Q}}_{h,\epsilon} \times \overline{\mathcal{Q}}_{h,\epsilon}$ such that

$$
\|\overline{Q}_{h+1} - \overline{Q}'_{h+1}\|_\infty \leqslant \epsilon, \qquad \|\underline{Q}_{h+1} - \underline{Q}'_{h+1}\|_\infty \leqslant \epsilon,
$$

where $\epsilon = 1/KH$. Let $V'_{h+1}(s)$ be the NE value of payoff matrix $\overline{Q}'_{h+1}(s,\cdot,\cdot)$. By Lemma E.2 we have

$$\left|V'_{h+1}(s) - \overline{V}_{h+1}(s)\right| \leqslant \epsilon.$$

Then we obtain

$$\begin{aligned}
|(\text{ii})| &= \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^K \left(\epsilon_h^\tau(\overline{V}_{h+1}) - \epsilon_h^\tau(V)\right)\phi_h^\tau + \phi^\top \Lambda_h^{-1} \sum_{\tau=1}^K \epsilon_h^\tau(\overline{V}'_{h+1})\phi_h^\tau\right| \\
&\leqslant \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^K \left(\epsilon_h^\tau(\overline{V}_{h+1}) - \epsilon_h^\tau(\overline{V}'_{h+1})\right)\phi_h^\tau\right| + \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^K \epsilon_h^\tau(\overline{V}'_{h+1})\phi_h^\tau\right| \\
&\leqslant 2\epsilon K \|\phi\|_{\Lambda_h^{-1}} + \left|\phi^\top \Lambda_h^{-1} \sum_{\tau=1}^K \epsilon_h^\tau(\overline{V}'_{h+1})\phi_h^\tau\right| \qquad \text{(A.10)} \\
&\leqslant 2\epsilon K \|\phi\|_{\Lambda_h^{-1}} + \underbrace{\left\|\sum_{\tau=1}^K \epsilon_h^\tau(\overline{V}'_{h+1})\phi_h^\tau\right\|_{\Lambda_h^{-1}}}_{\text{(iii)}} \|\phi\|_{\Lambda_h^{-1}}. \qquad \text{(A.11)}
\end{aligned}$$

For any $\tau \in [K], h \in [H]$, we define

$$\mathcal{F}_{h,\tau-1} := \sigma\left(\{(s_h^j, a_h^j, b_h^j)\}_{j=1}^{\min\{\tau+1,K\}} \cup \{(r_h^j, s_{h+1}^j)\}_{j=1}^{\tau}\right),$$

where $\sigma(\cdot)$ is the $\sigma-$algebra generated by a set of random variables. For all $\tau \in [K]$, we have $\phi(s_h^\tau, a_h^\tau, b_h^\tau) \in \mathcal{F}_{h,\tau-1}$, as $(s_h^\tau, a_h^\tau, b_h^\tau)$ is $\mathcal{F}_{h,\tau-1}-$measurable. Besides, for any fix function $V : \mathcal{S} \to [0, H-1]$ and all $\tau \in [K]$, we have

$$\epsilon_h^\tau(V) = r_h^\tau + V(s_{h+1}^\tau) - (\mathbb{B}_h V)(s_h^\tau, a_h^\tau, b_h^\tau) \in \mathcal{F}_{h,\tau-1}$$

and $\{\epsilon_h^\tau(V)\}_{\tau=1}^K$ is a stochastic process adapted to the filtration $\{\mathcal{F}_{h,\tau}\}_{\tau=0}^K$. By Lemma E.4, we obtain an estimation of term (iii). For any $\delta \in (0,1)$,

$$\mathbb{P}\left(\left\|\sum_{\tau=1}^K \epsilon_h^\tau(V)\phi_h^\tau\right\|_{\Lambda_h^{-1}}^2 \geqslant 2H^2 \log\left(\frac{\det(\Lambda_h)^{1/2}}{\delta \det(I)^{1/2}}\right)\right) \leqslant \delta$$

Since

$$\|\Lambda_h\|_{\text{op}} = \left\|I + \sum_{\tau=1}^K \phi_h^\tau (\phi_h^\tau)^\top\right\|_{\text{op}} \leqslant 1 + \sum_{\tau=1}^K \left\|\phi_h^\tau (\phi_h^\tau)^\top\right\|_{\text{op}} \leqslant 1 + K,$$

we have $\det(\Lambda_h) \leqslant (1+K)^d$, which further implies

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{\tau=1}^K \epsilon_h^\tau(V)\phi_h^\tau\right\|_{\Lambda_h^{-1}}^2 \geqslant H^2\left(d\log(1+K) + 2\log(\frac{1}{\delta})\right)\right) \\
&\leqslant \mathbb{P}\left(\left\|\sum_{\tau=1}^K \epsilon_h^\tau(V)\phi_h^\tau\right\|_{\Lambda_h^{-1}}^2 \geqslant 2H^2 \log\left(\frac{\det(\Lambda_h)^{1/2}}{\delta \det(I)^{1/2}}\right)\right) \leqslant \delta.
\end{aligned}$$

By Lemma E.3, $\left|\underline{\mathcal{Q}}_{\epsilon,h} \times \overline{\mathcal{Q}}_{\epsilon,h}\right| = \mathcal{N}_{h,\epsilon}^2 \leqslant \left(1 + \frac{4H\sqrt{dK}}{\epsilon}\right)^{2d}\left(1 + \frac{8\beta^2\sqrt{d}}{\epsilon^2}\right)^{2d^2}$. Thus, by the union bound argument we have

$$|(\text{iii})| \lesssim dH\sqrt{\zeta}\|\phi\|_{\Lambda_h^{-1}}. \qquad \text{(A.12)}$$

with probability at least $1 - p/2$. Combining (A.8), (A.10), and (A.12), we have

$$\left|\phi^\top \overline{w}_h - (\mathbb{B}_h \overline{V}_{h+1})(s,a,b)\right| \leq \beta\|\phi\|_{\Lambda_h^{-1}} = \Gamma_h(s,a,b)$$

with probability at least $1 - p/2$. Then, we have

$$\phi^\top \overline{w}_h + \Gamma_h(s,a,b) \geq \mathbb{B}_h \overline{V}_{h+1}(s,a,b) \geq -(H-h+1).$$

The last inequality follows from $|r_h| \leqslant 1$ and $\left|\overline{V}_{h+1}(s,a,b)\right| \leqslant H-h$. The inequality implies

$$\overline{Q}_h(s,a,b) = \min\{H-h+1, \phi^\top \overline{w}_h + \Gamma_h(s,a,b)\} \leq \phi^\top \overline{w}_h + \Gamma_h(s,a,b).$$

Therefore, we have

$$\begin{aligned}\bar{\iota}_h(s_h,a_h,b_h) &= \mathbb{B}_h \overline{V}_{h+1}(s_h,a_h,b_h) - \overline{Q}_h(s_h,a_h,b_h) \\ &\geq \mathbb{B}_h \overline{V}_{h+1}(s_h,a_h,b_h) - \phi^\top \overline{w}_h - \Gamma_h(s_h,a_h,b_h) \geq -2\Gamma_h(s_h,a_h,b_h). \end{aligned} \tag{A.13}$$

If $\phi^\top \overline{w}_h + \Gamma_h(s,a,b) \geq H-h+1$, then, we have

$$\overline{Q}_h(s,a,b) = \min\{H-h+1, \phi^\top \overline{w}_h + \Gamma_h(s,a,b)\} = H-h+1.$$

Thus, we further obtain that

$$\bar{\iota}_h(s,a,b) = \mathbb{B}_h \overline{V}_{h+1}(s,a,b) - \overline{Q}_h(s,a,b) = \mathbb{B}_h \overline{V}_{h+1}(s,a,b) - (H-h+1) \leq 0. \tag{A.14}$$

Otherwise, $\phi^\top \overline{w}_h + \Gamma_h(s,a,b) \leq H-h+1$, which implies $\overline{Q}_h(s,a,b) = \phi^\top \overline{w}_h + \Gamma_h(s,a,b)$. In this situation, we have

$$\begin{aligned}\bar{\iota}_h(s,a,b) &= \mathbb{B}_h \overline{V}_{h+1}(s,a,b) - \overline{Q}_h(s,a,b) \\ &= \mathbb{B}_h \overline{V}_{h+1}(s,a,b) - \phi^\top \overline{w}_h - \Gamma_h(s,a,b) \leq 0. \end{aligned} \tag{A.15}$$

Similarly, we can prove

$$0 \leq \underline{\iota}_h(s,a,b) \leq 2\Gamma_h(s,a,b) \tag{A.16}$$

with probability at least $1 - p/2$. Thus, the event $\mathcal{E}$ happens with probability at least $1-p$, which concludes our proof. □

A.2. Proof of Lemma A.2

Proof of Lemma A.2. We prove the first inequality i.e.,

$$\underline{V}_h(x) \leq V_h^{\widehat{\pi},*}(x).$$

We prove it by induction. When $h = H+1, V_h^{\widehat{\pi},*} = \underline{V}_h = 0$, the inequality holds trivially. Now we suppose the inequality holds for step $h+1$, we prove it also holds for step h. By definition of value function,

$$\begin{aligned} V_h^{\widehat{\pi},*}(x) - \underline{V}_h(x) &= \mathbb{E}_{\widehat{\pi},*}[Q_h^{\widehat{\pi},*}(x,a,b)] - \mathbb{E}_{\widehat{\pi},\nu'}[\underline{Q}_h(x,a,b)] \\ &= \mathbb{E}_{\widehat{\pi},*}[Q_h^{\widehat{\pi},*}(x,a,b) - \underline{Q}_h(x,a,b)] \\ &\quad + \big(\mathbb{E}_{\widehat{\pi},*}[\underline{Q}_h(x,a,b)] - \mathbb{E}_{\widehat{\pi},\nu'}[\underline{Q}_h(x,a,b)]\big). \end{aligned} \tag{A.17}$$

By the definition that $\underline{\iota}_h(x,a,b) = \mathbb{B}_h \underline{V}_{h+1}(x,a,b) - \underline{Q}_h(x,a,b)$, we have

$$Q_h^{\widehat{\pi},*}(x,a,b) - \underline{Q}_h(x,a,b) = \mathbb{B}_h\big(V_{h+1}^{\widehat{\pi},*}(x,a,b) - \underline{V}_{h+1}(x,a,b)\big) + \underline{\iota}_h(x,a,b) \geq 0, \tag{A.18}$$

where the last inequality follows from Lemma A.1 and induction assumption. Meanwhile, by the property of NE, we have

$$\mathbb{E}_{\widehat{\pi},*}[\underline{Q}_h(x,a,b)] - \mathbb{E}_{\widehat{\pi},\nu'}[\underline{Q}_h(x,a,b)] \geq 0, \tag{A.19}$$

Combining (A.17), (A.18) and (A.19), we obtain

$$V_h^{\widehat{\pi},*}(x) - \underline{V}_h(x) \geq 0,$$

which concludes the proof. □

A.3. Proof of Lemma A.4

Proof of Lemma A.4. We estimate the each term in the summation. By the fact that $(\pi'_h(\cdot|x), \widehat{\nu}_h(\cdot|x))$ is the NE of the matrix game with payoff $\overline{Q}_h(x,\cdot,\cdot)$ for any $x \in \mathcal{S}$, we have

$$\langle \overline{Q}_h(s_h,\cdot,\cdot), \pi'_h(\cdot|s_h) \otimes \widehat{\nu}_h(\cdot|s_h) - \pi'_h(\cdot|s_h) \otimes \nu^*_h(\cdot|s_h) \leq 0. \tag{A.20}$$

Taking summation over $h \in [H]$, we obtain

$$\sum_{h=1}^{H} \mathbb{E}_{\pi',\nu^*} \Big[\langle \overline{Q}_h(s_h,\cdot,\cdot), \pi'_h(\cdot|s_h) \otimes \widehat{\nu}_h(\cdot|s_h) - \pi'_h(\cdot|s_h) \otimes \nu^*_h(\cdot|s_h) \rangle | s_1 = x \Big] \leq 0,$$

which concludes the proof. □

B. Proof of Corollary 4.3

Proof of Corollary 4.3. Fix (π^*, ν). For notational simplicity, we define

$$\Sigma_h(x) = \mathbb{E}_{\pi^*,\nu}[\phi(s_h,a_h,b_h)\phi^\top(s_h,a_h,b_h)|s_1 = x].$$

By the assumption, we have

$$\begin{aligned}
\mathbb{E}_{\pi^*,\nu}\left[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}\right] &\leqslant \mathbb{E}_{\pi^*,\nu}\left[\sqrt{\phi_h^\top (I + c_1 K \Sigma_h(x))^{-1}\phi_h}|s_1 = x\right] \\
&= \mathbb{E}_{\pi^*,\nu}\left[\sqrt{\mathrm{tr}((I + c_1 K \Sigma_h(x))^{-1}\phi_h\phi_h^\top)}|s_1 = x\right] \\
&\leqslant \sqrt{\mathbb{E}_{\pi^*,\nu}\left[\mathrm{tr}((I + c_1 K \Sigma_h(x))^{-1}\phi_h\phi_h^\top)|s_1 = x\right]} \\
&= \sqrt{[\mathrm{tr}((I + c_1 K \Sigma_h(x))^{-1}\Sigma_h(x))]} \\
&= \sqrt{\frac{1}{c_1 K}} \cdot \sqrt{\mathrm{tr}((I + c_1 K \Sigma_h(x))^{-1}(c_1 K \Sigma_h(x) + I - I))} \\
&= \sqrt{\frac{1}{c_1 K}} \cdot \sqrt{\mathrm{tr}(I - (I + c_1 K \Sigma_h(x))^{-1})} \leqslant \sqrt{\frac{d}{c_1 K}},
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality. Thus, for any policy ν, we have

$$2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu}[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}|s_1 = x] \leqslant 2cc_1^{-1/2} d^{3/2} H^2 K^{-1/2}\sqrt{\zeta}.$$

Similarly, for any policy π, we have

$$2\beta \sum_{h=1}^{H} \mathbb{E}_{\pi,\nu^*}[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}|s_1 = x] \leqslant 2cc_1^{-1/2} d^{3/2} H^2 K^{-1/2}\sqrt{\zeta}.$$

Let $c' = 4cc_1^{-1/2}$, by the definition of related uncertainty, we obtain

$$4\beta \cdot \mathrm{RU}(\mathcal{D}, x) \leqslant c' d^{3/2} H^2 K^{-1/2}\sqrt{\zeta}.$$

Combined with Theorem 3.3, we further obtain

$$\mathrm{SubOpt}(\mathrm{PMVI}(\mathcal{D}), x) \leqslant c' d^{3/2} H^2 K^{-1/2}\sqrt{\zeta},$$

which concludes our proof. □

C. Proof of Corollary 4.4

Proof of Corollary 4.4. The proof consists of two steps. In the first step, we use Lemma E.5 for concentration. In the second step, we estimate the suboptimality. Recall that we denote $\phi_h = \phi(s_h, a_h, b_h)$, $\phi_h^\tau = \phi(s_h^\tau, a_h^\tau, b_h^\tau)$ and $\Sigma_h(x) = \mathbb{E}_{\overline{\pi},\overline{\nu}}[\phi(s_h, a_h, b_h)\phi^\top(s_h, a_h, b_h)|s_1 = x]$. Let

$$Z_h = \sum_{\tau=1}^{K} A_h^\tau, \quad A_h^\tau = \phi_h^\tau(\phi_h^\tau)^\top - \Sigma_h, \quad \forall h \in [H].$$

Clearly, we have $\mathbb{E}_{\overline{\pi},\overline{\nu}}[A_h^\tau] = 0$. Since the trajectories are induced by the behavior policy $(\overline{\pi}, \overline{\nu})$, the K trajectory are i.i.d. and $\{A_h^\tau\}$ are i.i.d.. By Assumption 2.1, we have $\|\phi(s, a, b)\| \leqslant 1$ for any $(s, a, b) \in \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2$, which further implies $\left\|\phi_h^\tau(\phi_h^\tau)^\top\right\|_{\text{op}} \leqslant 1$. Then, we have

$$\|\Sigma_h\|_{\text{op}} = \left\|\mathbb{E}_{\overline{\pi},\overline{\nu}}[\phi_h^\tau(\phi_h^\tau)^\top]\right\|_{\text{op}} \leqslant \mathbb{E}_{\overline{\pi},\overline{\nu}}\left[\left\|\phi_h^\tau(\phi_h^\tau)^\top\right\|_{\text{op}}\right] \leqslant 1.$$

Thus, we obtain

$$\|A_h^\tau\|_{\text{op}} \leqslant \|\Sigma_h\|_{\text{op}} + \left\|\phi_h^\tau(\phi_h^\tau)^\top\right\|_{\text{op}} \leqslant 2,$$

and

$$\left\|A_h^\tau(A_h^\tau)^\top\right\|_{\text{op}} \leqslant \|A_h^\tau\|_{\text{op}}^2 \leqslant 4.$$

Since $\{A_h^\tau\}_{\tau\in[K}$ are i.i.d. and mean-zero, for any $h \in [H]$, we have

$$\left\|\mathbb{E}_{\overline{\pi},\overline{\nu}}[Z_h^\top Z_h]\right\|_{\text{op}} = \left\|\sum_{\tau=1}^{K} \mathbb{E}_{\overline{\pi},\overline{\nu}}[A_h^\tau(A_h^\tau)^\top]\right\|_{\text{op}} = K \cdot \left\|\mathbb{E}_{\overline{\pi},\overline{\nu}}[A_h^\tau(A_h^\tau)^\top]\right\|_{\text{op}} \leqslant 4K.$$

Similarly, we can obtain $\left\|\mathbb{E}_{\overline{\pi},\overline{\nu}}[Z_h Z_h^\top]\right\|_{\text{op}} \leqslant 4K$. By Lemma E.5, we have ,

$$\mathbb{P}(\|Z_h\|_{\text{op}} \geqslant t) \leqslant 2d \cdot \exp\left(-\frac{t^2/2}{4K + 2t/3}\right), \quad \forall t > 0.$$

Let $t = \sqrt{10K \log(4dH/p)}$, we have

$$\mathbb{P}(\|Z_h\|_{\text{op}} \geqslant t) \leqslant \frac{p}{2H}.$$

By the definition of Λ_h, we have

$$Z_h = \sum_{\tau=1}^{K} \phi_h^\tau(\phi_h^\tau)^\top - K\Sigma_h = \Lambda_h - I - K\Sigma_h.$$

When $K > 40/\underline{c} \cdot \log(4dH/p)$, it holds with probability at least $1 - p/2$ that

$$\begin{aligned}
\lambda_{\min}(\Lambda_h) &= \lambda_{\min}(Z_h + I + K\Sigma_h) \\
&\geqslant K\lambda_{\min}(\Sigma_h) - \|Z_h\|_{\text{op}} \geqslant K\left(\underline{c} - \sqrt{10/K \cdot \log(4dH/p)}\right) \\
&\geqslant K\underline{c}/2
\end{aligned}$$

for all $h \in [H]$. Let $c'' = \sqrt{2/\underline{c}}$, with probability $1 - p/2$, we have

$$\left\|\Lambda_h^{-1}\right\|_{\text{op}} \leqslant c''^2 K^{-1}$$

for all $h \in [H]$. Combined the fact that $\|\phi\|(\cdot, \cdot, \cdot) \leqslant 1$, for all $h \in [H]$, we have

$$\sqrt{\phi_h^\top \Lambda_h^{-1} \phi_h} \leq \left\|\Lambda_h^{-1}\right\|_{\text{op}} \leqslant c'' K^{-1/2}.$$

Then, for any policy pair (π, ν), we have

$$\sum_{h=1}^{H} \mathbb{E}_{\pi,\nu}\left[\sqrt{\phi_h^\top \Lambda_h^{-1}\phi_h}|s_1 = x\right] \leqslant c'' H K^{-1/2}.$$

Together with Theorem 3.3, we have

$$\text{SubOpt}(\text{PMVI}(\mathcal{D}), x) \leqslant c' d H^2 K^{-1/2}\sqrt{\zeta}$$

with probability at least $1-p$ with $c' = 4cc''$. Therefore, we finish the proof. □

D. Proof of the Information-Theoretic Lower Bound

The proof is organized as follows. First, we construct a class of linear MGs $\mathfrak{M}$ and a dataset collecting process for $\mathcal{D}$ which is compliant with the underlying MG. Then, given a policy pair (π, ν) constructed based on $\mathcal{D}$, we find two hard MGs $\mathcal{M}_1, \mathcal{M}_2$ from the class $\mathfrak{M}$ such that the policy pair cannot achieve a desired suboptimality simultaneously. Before continuing, we introduce another notion of suboptimality, defined as

$$\text{SubOpt}_{\text{w}}((\pi,\nu), x_0) = |V_1^* - V_1^{\pi,\nu}| \leq V_1^{*,\nu} - V_1^{\pi,*} = \text{SubOpt}((\pi,\nu), x_0),$$

due to the weak duality property given in (2.5). Therefore, we can prove the lower bound for $\text{SubOpt}_{\text{w}}((\pi,\nu), x_0)$ which implies the original theorem.

D.1. Construction of the Linear MG Class $\mathfrak{M}$

The class $\mathfrak{M}$ is defined to be

$$\mathfrak{M} = \{M(p_1, p_2, p_3) : p_1, p_2 \in [1/4, 3/4], p_3 = \min\{p_1, p_2\}\},$$

where $M(p_1, p_2, p_3)$ is a MG with $H \geqslant 2$, state space $\mathcal{S} = \{x_0, x_1, x_2\}$ and action space $|\mathcal{A}_1| = |\mathcal{A}_2| = \{y_i\}_{i=0}^{A}$ with $A = |\mathcal{A}_1| \geq 3$. We fix the initial state as x_0. Now we define the transition kernel of the game at step $h = 1$ to be

$$\begin{aligned}
&\mathbb{P}_1(x_1 \mid x_0, y_1, y_j) = p_1 \quad \mathbb{P}_1(x_2 \mid x_0, y_1, y_j) = 1 - p_1,\ \forall j \in \{1, \cdots, A\},\\
&\mathbb{P}_1(x_1 \mid x_0, y_2, y_j) = p_2, \quad \mathbb{P}_1(x_2 \mid x_0, y_2, y_j) = 1 - p_2,\ \forall j \in \{1, \cdots, A\},\\
&\mathbb{P}_1(x_1 \mid x_0, y_i, y_j) = p_3, \quad \mathbb{P}_1(x_2 \mid x_0, y_i, y_j) = 1 - p_3,\ \forall i \geq 3, \forall j \in \{1, \cdots, A\}.
\end{aligned}$$

According to the construction, we can see that the transition is determined by the max-player's action at step $h = 1$. At subsequent step $h \geq 2$, we set

$$\mathbb{P}_h(x_1|x_1, y_i, y_j) = \mathbb{P}_h(x_2|x_2, y_i, y_j) = 1, \forall i, j \in \{1, \cdots, A\}.$$

In other words, the states x_1 and x_2 are absorbing. The reward functions of the game are defined as

$$\begin{aligned}
&r_1(x_0, y_i, y_j) = 0,\ \forall i, j \in \{1, \cdots, A\},\\
&r_h(x_1, y_i, y_j) = 1, r_h(x_2, y_i, y_j) = 0,\ \forall h \geq 2, \forall i, j \in \{1, \cdots, A\}.
\end{aligned}$$

We further illustrate the class $\mathfrak{M}$ in Figure D.1. To show that the game $\mathcal{M}(p_1, p_2, p_3)$ is indeed a linear MG, we define the feature map $\phi(s, a, b)$ to be

$$\begin{aligned}
&\phi(x_0, y_i, y_j) = (e_{(i-1)A+j}, 0, 0) \in \mathbb{R}^{A^2+2} \quad \phi(x_1, y_i, y_j) = (\mathbf{0}_{A^2}, 1, 0) \in \mathbb{R}^{A^2+2}\\
&\phi(x_2, y_i, y_j) = (\mathbf{0}_{A^2}, 0, 1) \in \mathbb{R}^{A^2+2} \quad \forall i, j \in \{1, \cdots, A\},
\end{aligned}$$

where $e_n \in \mathbb{R}^{A^2}$ is a vector whose components are all zero except for the n-th one.

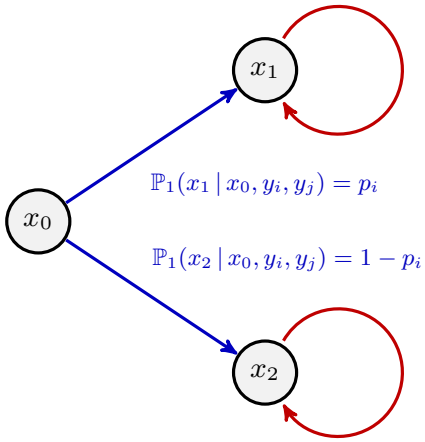

Figure 1. Illustration of the Game $\mathcal{M}(p_1, p_2, p_3)$: In the first step with initial state x_0, the game is totally determined by the max-player's action. The game has a probability of p_i to enter state x_1 if the max-player takes action $a_1 = y_i$. Meanwhile, x_1 and x_2 are absorbing states.

D.2. Dataset Collecting Process

We speficy the dataset collecting process in this subsection. Given an MG $\mathcal{M}(p_1, p_2, p_3) \in \mathfrak{M}$, the dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, b_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ consists of K trajectories starting from the initial state x_0, namely, $x_1^\tau = x_0$ for all $\tau \in [K]$. The actions taken at the first step $\{a_1^\tau, b_1^\tau\}_{\tau=1}^K$ are predetermined. The transitions at step $h = 1$ are sampled from $\mathcal{M}$ and are independent across K trajectories. The rewards are also generated by the $\mathcal{M}$. The subsequent actions $\{a_h^\tau, b_h^\tau\}_{\tau=1}^K, h \geq 2$ are arbitrary since they do not affect the transition and reward generation. In this case, the dataset is compliant with the underlying MG $\mathcal{M}$.

Before continuing, we define

$$\begin{aligned} n_{ij} &= \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = y_i, b_1^\tau = y_j\}, \quad \kappa_i^j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = y_i, s_2^\tau = x_j\}, \\ n_{\min}^k &= \min\{\min_j n_{kj}, \min_i n_{ik}\}, \quad n_i = \sum_{j=1}^A n_{ij}, \quad m_i = \sum_{\tau=1}^K \mathbb{1}\{s_2^\tau = x_i\}. \end{aligned} \tag{D.1}$$

In other words, in the dataset $\mathcal{D}$, the action pair (y_i, y_j) is taken by two players at step $h = 1$ for n_{ij} times; the event that the max-player takes action y_i at step $h = 1$ and the next state is x_j happens for κ_i^j times; the max-player takes action y_i at step $h = 1$ for n_i times; and the initial state x_0 transits to x_i for m_i times. Finally, $n_{\min}^k$ measures how well the dataset $\mathcal{D}$ covers the state action pairs where one action is fixed to be k.

Since x_1 and x_2 are absorbing states, for learning the optimal policy π^*, the original dataset $\mathcal{D}$ contains the same information as the reduced one $\mathcal{D}_1 := \{(x_1^\tau, a_1^\tau, b_1^\tau, x_2^\tau, r_2^\tau)\}_{\tau=1}^K$. Recall that the actions at step $h = 1$ are predetermined. The randomness of the dataset generation only comes from the transiton at the first step and we can write:

$$\begin{aligned} \mathbb{P}_{\mathcal{D}\sim\mathcal{M}}(\mathcal{D}_1) &= \prod_{\tau=1}^K \mathbb{P}_{\mathcal{M}}(s_2 = x_2^\tau \mid s_1 = x_1^\tau = x_0, a_1 = a_1^\tau, b_1 = b_1^\tau) \\ &= \prod_{j=1}^A \left(p_j^{k_j^1}(1-p_j)^{k_j^2}\right). \end{aligned} \tag{D.2}$$

D.3. Lower Bound of the Suboptimality

In this subsection, we constructed two MGs and show that the suboptimality of any algorithm that outputs a policy based on the dataset $\mathcal{D}$ is lower bounded by the hypothesis testing risk and the risk can be further lower bounded by some tuning parameters.

Lemma D.1 (Reduction to Testing)**.** *For the dataset $\mathcal{D}$ collected as specified in Section D.2, there exists two MGs $\mathcal{M}_1(p^*, p, p), \mathcal{M}_2(p, p^*, p) \in \mathfrak{M}$ where $p^* > p$ satisfy $p, p^* \in [1/4, 3/4]$, such that the output policy $\mathrm{Algo}(\mathcal{D})$ of any algorithm satisfies:*

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_l}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right]\\ &\geqslant (H-1)(p^*-p)\left(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_l}\left[1-\widehat{\pi}_1(y_1)\right] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\widehat{\pi}_1(y_1)\right]\right),\end{aligned}$$

It further holds that

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_l}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right] \geqslant 1/2(H-1)(p^*-p). \tag{D.3}$$

The right-hand side of (D.3) is the risk of a (randomized) test function about the hypothesis testing problem:

$$H_0 : \mathcal{M} = \mathcal{M}_1 \text{ versus } H_1 : \mathcal{M} = \mathcal{M}_2.$$

This construction mirrors the Le Cam method (Le Cam, 2012; Yu et al., 1997). See the Section 5.3.2 of Jin et al. (2020b) for a detailed discussion.

Proof of Lemma D.1. We first notice that by the construction of $\mathcal{M}_1$ and $\mathcal{M}_2$, the games degrade to the MDPs. Therefore, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right] &= \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\left|V_1^{\pi^*,\nu^*}(x_0) - V_1^{\widehat{\pi},\widehat{\nu}}(x_0)\right|\right]\\ &= \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[V_1^{\pi^*,\nu^*}(x_0) - V_1^{\widehat{\pi},\widehat{\nu}}(x_0)\right],\end{aligned}$$

where we use the fact that the Nash value is the V-value of the induced MDP in the last equality. Clearly, for $\mathcal{M}_1$, π_1^* puts probability 1 for action y_1 given the state x_0. In this case, we have the following calculation:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[V_1^{\pi^*,\nu^*}(x_0) - V_1^{\widehat{\pi},\widehat{\nu}}(x_0)\right] &= (H-1)\left(p^* - \sum_{i=1}^{A}\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\widehat{\pi}_1(y_i)\right]p_i\right)\\ &= (H-1)\mathbf{E}_{\mathcal{D}\sim\mathcal{M}_1}\left(p^*(1-\widehat{\pi}_1(y_1)) - \sum_{i\neq 1}\widehat{\pi}_i(y_i)p\right)\\ &= (H-1)(p^*-p)\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[1-\widehat{\pi}_1(y_1)\right],\end{aligned}$$

where we use $\sum_{j=1}^{A}\widehat{\pi}_1(y_j) = 1$ in the last equality. Therefore, we have

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right] = (H-1)(p^*-p)\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[1-\widehat{\pi}_1(y_1)\right].$$

Similarly,

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right] = (H-1)(p^*-p)\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[1-\widehat{\pi}_1(y_2)\right].$$

It follows that

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}); x_0)\right]\\ &\quad\geqslant (H-1)(p^*-p)\left(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[1-\widehat{\pi}_1(y_1)\right] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\widehat{\pi}_1(y_1)\right]\right),\end{aligned} \tag{D.4}$$

where we use $\widehat{\pi}_1(y_1) \leq 1 - \widehat{\pi}_1(y_2)$. This concludes the proof of (D.3). It remains to find a lower bound for the right-hand side. We have:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[1-\widehat{\pi}_1(y_1)\right] + \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\widehat{\pi}_1(y_1)\right] &\geqslant 1 - \mathrm{TV}\left(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}, \mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}\right)\\ &\geqslant 1 - \sqrt{\mathrm{KL}\left(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}||\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}\right)/2}\end{aligned} \tag{D.5}$$

where $\mathrm{TV}(\cdot,\cdot)$ and $\mathrm{KL}(\cdot||\cdot)$ are the total variation distance of probability measures and Kullback-Leibler (KL) divergence of two distributions, respectively. Here the first inequality comes from the definition of total variation distance, and the last inequality follows from Pinsker's inequality. Intuitively, we set p and p^* carefully to make $\mathcal{M}_1$ and $\mathcal{M}_2$ hard to be distinguished.

As stated in (D.2), we can explicitly write down the probability of the reduced dataset $\mathcal{D}_1$ as

$$\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}(\mathcal{D}_1) = (p^*)^{\kappa_1^1}\cdot(1-p^*)^{\kappa_1^2}\cdot p^{\sum_{i\neq 1}\kappa_i^1}\cdot(1-p)^{\sum_{i\neq 1}\kappa_i^2};$$
$$\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}(\mathcal{D}_1) = (p^*)^{\kappa_2^1}\cdot(1-p^*)^{\kappa_2^2}\cdot p^{\sum_{i\neq 2}\kappa_i^1}\cdot(1-p)^{\sum_{i\neq 2}\kappa_i^2}.$$

We recall $\kappa_i^j = \sum_{\tau=1}^K \mathbb{1}\{a_1^\tau = y_i, s_2^\tau = x_j\}$. Since the randomness only comes from the state transition at the first step for $\mathcal{D}_1$ and these transitions are independent across K trajectories. It follows that

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}||\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}\right)\\
&\quad= \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\left(\kappa_1^1-\kappa_2^1\right)\log\left(\frac{p^*}{p}\right)+\left(\kappa_1^2-\kappa_2^2\right)\log\left(\frac{1-p^*}{1-p}\right)\right]\\
&\quad= \left(p^*n_1-pn_2\right)\log\left(\frac{p^*}{p}\right)+\left((1-p^*)n_1-(1-p)n_2\right)\log\left(\frac{1-p^*}{1-p}\right)\\
&\quad= n_1\left(p^*\log\frac{p^*}{p}+(1-p^*)\log\frac{1-p^*}{1-p}\right)+n_2\left(p\log\frac{p}{p^*}+(1-p)\log\frac{1-p}{1-p^*}\right).
\end{aligned}\tag{D.6}$$

It remains to carefully set p and p^* to obtain the desired lower bound. To this end, we set

$$p=\frac{1}{2}-\frac{1}{16}\sqrt{\frac{2}{n_1+n_2}},\quad p^*=\frac{1}{2}+\frac{1}{16}\sqrt{\frac{2}{n_1+n_2}},$$

such that $p, p^*\in[1/4,3/4]$ and

$$p^*-p<\frac{1}{4}<\min\{p,p^*,1-p^*,1-p\}.$$

As a result of the inequality $\log(1+x)\leqslant x, \forall x>-1$, we have

$$p^*\log\frac{p^*}{p}+(1-p^*)\log\frac{1-p^*}{1-p}\leqslant\frac{(p^*-p)^2}{p(1-p)}\leqslant 16(p^*-p)^2,$$
$$p\log\frac{p}{p^*}+(1-p)\log\frac{1-p}{1-p^*}\leqslant\frac{(p^*-p)^2}{p^*(1-p^*)}\leqslant 16(p^*-p)^2.$$

Thus,

$$\mathrm{KL}\left(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}||\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}\right)\leqslant 16n_1(p^*-p)^2+16n_2(p^*-p)^2\leqslant 16(n_1+n_2)(p^*-p)^2\leq\frac{1}{2}.$$

It follows that

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[1-\widehat{\pi}_1(y_1)\right]+\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\widehat{\pi}_1(y_1)\right]\geqslant 1-\sqrt{\mathrm{KL}\left(\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_1}||\mathbb{P}_{\mathcal{D}\sim\mathcal{M}_2}\right)/2}\geqslant\frac{1}{2}.$$

Combined this with (D.4) and (D.5), we conclude that

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D});x_0)\right]+\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D});x_0)\right]\geqslant\frac{1}{2}(H-1)(p^*-p).\tag{D.7}$$

Therefore, we conclude the proof. □

D.4. Upper Bound of RU$(\mathcal{D}, x_0)$

We recall that we are concerning about

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}}\left[\frac{\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D});x_0\right)}{\mathrm{RU}(\mathcal{D},x_0)}\right].$$

We still need to find an upper bound of $\mathrm{RU}(\mathcal{D},x_0)$ for the constructed linear MGs.

Lemma D.2 (Upper Bound of $\mathrm{RU}(\mathcal{D}, x_0)$)**.** *Suppose the Assumption 2.4 holds and the underlying MG is $\mathcal{M} \in \mathfrak{M}$. We define $j^* = \mathrm{argmax}_{j\in\{1,2\}} p_j$ (we assume that $p_1 \neq p_2$). Then, the optimal policy satisfies $\pi_1^*(y_{j^*}) = 1$ and we further take $\nu_1^*(y_{j^*}) = 1$. Then, for Algorithm 1, it holds that*

$$\begin{aligned}
&\sum_{h=1}^{H} \sup_{\nu} \mathbb{E}_{\pi^*,\nu}\left[\left(\phi\left(s_h,a_h,b_h\right)^\top \Lambda_h^{-1}\phi\left(s_h,a_h,b_h\right)\right)^{1/2} \mid x_0\right] \\
&\qquad \leq \frac{1}{\sqrt{1+n_{min}^{j^*}}} + (H-1)\cdot\left(\frac{p_{j^*}}{\sqrt{1+m_1}} + \frac{1-p_{j^*}}{\sqrt{1+m_2}}\right), \\
&\sum_{h=1}^{H} \sup_{\pi} \mathbb{E}_{\pi,\nu^*}\left[\left(\phi\left(s_h,a_h,b_h\right)^\top \Lambda_h^{-1}\phi\left(s_h,a_h,b_h\right)\right)^{1/2} \mid x_0\right] \\
&\qquad \leq \sup_{\pi}\left\{\frac{1}{\sqrt{1+n_{min}^{j^*}}} + \sum_{i\in[A]} \pi_1(y_i)(H-1)\cdot\left(\frac{p_i}{\sqrt{1+m_1}} + \frac{1-p_i}{\sqrt{1+m_2}}\right)\right\}.
\end{aligned} \tag{D.8}$$

Furthermore, with probability at least $1-\frac{1}{K}$, the following event holds

$$\overline{\mathcal{E}} = \{m_i \geqslant K/4 - \sqrt{2K\log(2K)} \mid i = 1,2\},$$

where the probability is taken with respect to $\mathbb{P}_{\mathcal{D}\sim\mathcal{M}}$. Under $\overline{\mathcal{E}}$, for $K \geqslant 32\log(2K)$, we can obtain

$$\mathrm{RU}(\mathcal{D},\pi^*,\nu^*,x_0) \leq \frac{1}{\sqrt{n_{min}^{j^*}}} + \frac{2\sqrt{2}(H-1)}{\sqrt{K}} \leq \frac{2\sqrt{2}H}{\sqrt{n_{min}^{j^*}}}. \tag{D.9}$$

Proof of Lemma D.2. **Proof of (D.8).** We first consider $\sup_\nu \mathbf{E}_{\pi^*,\nu}$. By $x_1^\tau = x_0$ for all $\tau \in [K]$ and the definition of Λ_h, we have

$$\begin{aligned}
\Lambda_1 &= I + \sum_{\tau=1}^{K} \phi\left(x_0,a_1^\tau,b_1^\tau\right)\phi\left(x_0,a_1^\tau,b_1^\tau\right)^\top \\
&= \mathrm{diag}\left(1+n_{11}, 1+n_{12}\ldots, 1+n_{AA}, 1, 1\right) \in \mathbb{R}^{(A^2+2)\times(A^2+2)},
\end{aligned}$$

where the second equality follows from the definition of ϕ. For $h \geq 2$, the state is x_1 or x_2, so we have

$$\Lambda_h = I + \sum_{\tau=1}^{K} \phi\left(x_h^\tau,a_h^\tau,b_h^\tau\right)\phi\left(x_h^\tau,a_h^\tau,b_h^\tau\right)^\top = \mathrm{diag}\left(1,1,\ldots,1,1+m_1,1+m_2\right) \in \mathbb{R}^{(A^2+2)\times(A^2+2)},$$

where the second equality follows from the definition of ϕ. Under (π^*,ν), we know that

$$\mathbb{P}_{\pi^*,\nu}(s_2 = x_1) = p_{j^*}, \quad \mathbb{P}_{\pi^*,\nu}(s_2 = x_2) = 1 - p_{j^*}.$$

It follows that

$$\begin{aligned}
&\sup_{\nu} \mathbb{E}_{\pi^*,\nu}\left[\left(\phi\left(s_h,a_h,b_h\right)^\top \Lambda_h^{-1}\phi\left(s_h,a_h,b_h\right)\right)^{1/2} \mid s_1 = x_0\right] \\
&\qquad \leq \begin{cases} \left(1+n_{\min}^{j^*}\right)^{-1/2}, & h = 1, \\ p_{j^*}\cdot(1+m_1)^{-1/2} + (1-p_{j^*})\cdot(1+m_2)^{-1/2}, & h \in \{2,\ldots,H\}, \end{cases}
\end{aligned}$$

where we use the definition of ϕ, and the fact that $n_{\min}^{j^*} \leq n_{j^*i}$ for all $i \in [A]$.

For $\sup_\pi \mathbf{E}_{\pi,\nu^*}$, the main difference lies in the distribution of s_2:

$$\mathbb{P}_{\pi,\nu^*}(s_2 = x_1) = \sum_{j\in[A]} \pi_1(y_j)p_j, \quad \mathbb{P}_{\pi,\nu^*}(s_2 = x_2) = 1 - \sum_{j\in[A]} \pi_1(y_j)p_j.$$

It follows that

$$\mathbb{E}_{\pi,\nu^*}\left[\left(\phi\left(s_h,a_h,b_h\right)^\top \Lambda_h^{-1}\phi\left(s_h,a_h,b_h\right)\right)^{1/2} \mid s_1=x_0\right]$$
$$\leq \begin{cases}\left(1+n_{\min}^{j^*}\right)^{-1/2}, & h=1,\\ \sum_{i\in[A]}\pi_1(y_i)(H-1)\cdot\left(\frac{p_i}{\sqrt{1+m_1}}+\frac{1-p_i}{\sqrt{1+m_2}}\right), & h\in\{2,\ldots,H\},\end{cases}$$

where we use the definition of ϕ, and the fact that $n_{\min}^{j^*}\leq n_{ij^*}$ for all $i\in[A]$. This concludes the proof of (D.8).

We now turn to the high-probability event:

$$\overline{\mathcal{E}}=\left\{m_i\geqslant K/4-\sqrt{\frac{1}{2}K\log(2K)}\mid i=1,2\right\}.$$

By construction, we know that $\frac{3}{4}\geq p_1,p_2\geq\frac{1}{4}$. Therefore, we know that $\mathbb{E}[m_i]\geqslant 1/4K$ for $i=1,2$. By Hoeffding's inequality, for any $\xi\in(0,1)$, with probability at least $1-\xi$, the following event happens

$$\left\{m_i\geqslant K/4-\sqrt{\frac{1}{2}K\log(2/\xi)}\mid i=1,2\right\}.$$

Setting $\xi=1/K$, we obtain the desired result.

Proof of (D.9). In particular, for $K\geqslant 32\log(2K)$, with probability at least $1-1/K$, we have

$$m_i\geqslant K/8,\ \ \forall i\in\{1,2\}.$$

The (D.9) follows directly from (D.8) and $m_i\geqslant K/8$. □

D.5. Proof of Theorem 5.1

We now invoke Lemma D.1 and Lemma D.2 to give a detailed proof of Theorem 5.1.

Proof of Theorem 5.1. Since the actions are predetermined, we can additionally assume that

$$\frac{1}{\overline{c}}\leqslant\frac{n_{\min}^1}{n_{\min}^2}\leqslant\overline{c},\quad \frac{1}{\overline{c}}\leqslant\frac{n_{2i}}{n_{\min}^2}\leqslant\overline{c},\ \ \frac{1}{\overline{c}}\leqslant\frac{n_{1i}}{n_{\min}^2}\leqslant\overline{c}\quad\forall i\in\{1,\cdots,n\},$$

where $\overline{c}$ is a positive constant. This assumption means that the numbers of action pairs whose components contain y_1 or y_2 are relatively uniform. By Lemma D.1, there exist two games $\mathcal{M}_1,\mathcal{M}_2$ such that

$$\begin{aligned}&\max_{i\in\{1,2\}}\sqrt{n_{\min}^i}\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_i}\left[\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}),x_0\right)\right]\\ &\quad\geqslant\frac{\sqrt{n_{\min}^1n_{\min}^2}}{\sqrt{n_{\min}^1}+\sqrt{n_{\min}^2}}\left(\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_1}\left[\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}),x_0\right)\right]+\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_2}\left[\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}),x_0\right)\right]\right)\\ &\quad\geqslant\frac{\sqrt{n_{\min}^1n_{\min}^2}}{\sqrt{n_{\min}^1}+\sqrt{n_{\min}^2}}\frac{1}{2}(H-1)(p^*-p)=\frac{\sqrt{2}}{16}\cdot(H-1)\cdot\frac{\sqrt{n_{\min}^1n_{\min}^2}}{\sqrt{n_{\min}^1}+\sqrt{n_{\min}^2}}\cdot\frac{1}{\sqrt{n_1+n_2}},\end{aligned}$$

where the first inequality is because $\max\{x,y\}\geq ax+(1-a)y$ for all $a\in[0,1]$ and $x,y\geq 0$ and the second inequality follows from Lemma D.1. Note that

$$n_1+n_2=\sum_{i=1}^{A}(n_{1i}+n_{2i})\leqslant 2\overline{c}An_{\min}^2.$$

Therefore, we have

$$\max_{i\in\{1,2\}}\sqrt{n_{\min}^i}\mathbb{E}_{\mathcal{D}\sim\mathcal{M}_i}\left[\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}),x_0\right)\right]\geqslant\frac{1}{16\sqrt{\overline{c}A}}\cdot(H-1)\cdot\frac{\sqrt{n_{\min}^1/n_{\min}^2}}{\sqrt{n_{\min}^1/n_{\min}^2}+1}\geqslant C,\tag{D.10}$$

where $C = \frac{1}{16\sqrt{\bar{c}A}} \cdot (H-1) \cdot \frac{1}{1+\sqrt{\bar{c}}}$. Here the last inequality is because $f(t) = \frac{t}{1+t}$ is increasing for $t \geq 0$. We now take the optimal policy of game $\mathcal{M}_i$ to be $\pi_1^*(y_i) = \nu_1^*(y_i) = 1$. It follows that

$$
\begin{aligned}
&\max_{i\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_i}\left[\frac{\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}), x_0\right)}{\mathrm{RU}(\mathcal{D}, x_0)}\right] \\
&\quad\geqslant \max_{i\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_i}\left[\frac{\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}), x_0\right)}{\mathrm{RU}(\mathcal{D}, x_0)}\mathbb{1}_{\overline{\mathcal{E}}}\right] \\
&\quad\geqslant \max_{i\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_i}\left[\frac{1}{C_1}\sqrt{n^i_{\min}}\,\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}), x_0)\mathbb{1}_{\overline{\mathcal{E}}}\right] \\
&\quad= \max_{i\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_i}\left[\frac{1}{C_1}\sqrt{n^i_{\min}}\,\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}), x_0\right)\right] \\
&\qquad - \mathbb{E}_{\mathcal{D}\sim\mathcal{M}_i}\left[\frac{1}{C_1}\sqrt{n^i_{\min}}\,\mathrm{SubOpt}_{\mathrm{w}}\left(\mathrm{Algo}(\mathcal{D}), x_0\right)\cdot\mathbb{1}_{\overline{\mathcal{E}}^c}\right] \\
&\quad\geqslant \frac{C}{C_1} - \frac{\sqrt{K}}{C_1}\cdot 2H\cdot\frac{1}{K} \\
&\quad\geqslant \frac{C}{2C_1} := C' > 0,
\end{aligned}
$$

where $C' = \frac{C}{2C_1}, C_1 = 2\sqrt{2}H$, and $C = \frac{1}{16\sqrt{\bar{c}A}} \cdot (H-1) \cdot \frac{1}{1+\sqrt{\bar{c}}}$. The second inequality follows from (D.9). The third inequality is because (D.10), $\mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}), x_0) \leq 2H$, $n^i_{\min} \leq K$, and $\mathbb{P}(\overline{\mathcal{E}}) \leq \frac{1}{K}$. The forth inequality holds for $K \geqslant \frac{16H^2}{C^2}$. By $\mathrm{SubOpt}(\mathrm{Algo}(\mathcal{D}), x_0) \geq \mathrm{SubOpt}_{\mathrm{w}}(\mathrm{Algo}(\mathcal{D}), x_0)$, we conclude the proof of Theorem 5.1. □

E. Technical Lemmas

Recall that we use shorthands

$$
\phi_h = \phi(s_h, a_h, b_h), \qquad \phi_h^\tau = \phi(s_h^\tau, a_h^\tau, b_h^\tau), \qquad r_h^\tau = r(s_h^\tau, a_h^\tau, b_h^\tau).
$$

Lemma E.1. *For any dataset $\mathcal{D}$, we define*

$$
w_h = \theta_h + \int_{x\in\mathcal{S}} \overline{V}_{h+1}(x)\mu_h(x)dx,
$$

where $\overline{V}_{h+1}(x)$ are the value function constructed in Algorithm 1, then w_h as well as $\overline{w}_h$ in Algorithm 1 satisfy

$$
\|w_h\| \leqslant H\sqrt{d}, \quad \|\overline{w}_h\| \leqslant H\sqrt{Kd}, \quad i = 1, 2.
$$

Proof of Lemma E.1. By definition of w_h,

$$
\begin{aligned}
\|w_h\| &= \left\|\theta_h + \int_{x\in\mathcal{S}} \overline{V}_{h+1}(x)\mu_h(x)\mathrm{d}x\right\| \leqslant \|\theta_h\| + \left\|\int_{x\in\mathcal{S}} \overline{V}_{h+1}(x)\mu_h(x)\mathrm{d}x\right\| \\
&\leqslant \sqrt{d} + (H-h)\int_{x\in\mathcal{S}} \|\mu_h(x)\|\,\mathrm{d}x \leqslant \sqrt{d} + (H-h)\sqrt{d} \leqslant H\sqrt{d}
\end{aligned}
$$

where the second and the last inequities follow from the regulation assumption in Assumption 2.1 that $\|\theta_h\| \leqslant \sqrt{d}$ and $\int_{x\in\mathcal{S}} \|\mu_h(x)\|\,\mathrm{d}x \leqslant \sqrt{d}$, while the construction in Algorithm 1 guarantees $\overline{V}_{h+1}(x) \leqslant H - h$, which implies the third inequality.

By construction of $\overline{w}_h$ in Algorithm 1,

$$
\|\overline{w}_h\| = \left\|\Lambda_h^{-1}\sum_{\tau=1}^K \phi_h^\tau(r_h^\tau + \overline{V}_{h+1}(s_{h+1}))\right\| \leqslant H\sum_{\tau=1}^K \left\|\Lambda_h^{-1}\phi_h^\tau\right\|,
$$

where the last inequality follows from triangle inequality and $|r_h^\tau| \leqslant 1, \overline{V}_{h+1}(x) \leqslant H - h$. Note that

$$\left\|\Lambda_h^{-1}\phi_h^\tau\right\| = \sqrt{(\phi_h^\tau)^\top \Lambda_h^{-1/2}\Lambda_h^{-1}\Lambda_h^{-1/2}\phi_h^\tau} \leqslant \left((\phi_h^\tau)^\top \Lambda_h^{-1}\phi_h^\tau\right)^{1/2}.$$

The last inequality follows from $\left\|\Lambda_h^{-1}\right\|_{\text{op}} \leqslant 1$. Thus,

$$\begin{aligned} H\sum_{\tau=1}^K \left\|\Lambda_h^{-1}\phi_h^\tau\right\| &= H\sum_{\tau=1}^K \left((\phi_h^\tau)^\top \Lambda_h^{-1}\phi_h^\tau\right)^{1/2} \leqslant H\sqrt{K}\left(\sum_{\tau=1}^K (\phi_h^\tau)^\top \Lambda_h^{-1}\phi_h^\tau\right)^{1/2} \\ &= H\sqrt{K}\left(\operatorname{tr}(\Lambda_h^{-1}\sum_{\tau=1}^K \phi_h^\tau(\phi_h^\tau)^\top)\right)^{1/2} = H\sqrt{K}\left(\operatorname{tr}(\Lambda_h^{-1}(\Lambda_h - 1I))\right)^{1/2} \\ &\leqslant H\sqrt{K}\left(\operatorname{tr}(I)\right)^{1/2} = H\sqrt{Kd}, \end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality. □

Lemma E.2 (Non-Expansive Property of Nash Value)**.** *For any integer n and matrix $A \in \mathbb{R}^{n\times n}$, we denote $f(A) = \max_{x\in\Delta}\min_{y\in\Delta} x^\mathsf{T} A y$, where $\Delta = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$. Fix $\epsilon > 0$, given any matrices $A_1, A_2 \in \mathbb{R}^{n\times n}$ satisfying $\|A_1 - A_2\|_\infty \leq \epsilon$, we have*

$$|f(A_1) - f(A_2)| \leq \epsilon.$$

Proof. Fix x, we have

$$\min_{y\in\Delta} x^\mathsf{T} A_1 y = \min_{y\in\Delta} x^\mathsf{T}(A_1 - A_2 + A_2)y \geq \min_{y\in\Delta} x^\mathsf{T} A_2 y - \epsilon,$$

where the inequality follows from the fact that $\|A_1 - A_2\|_\infty \leq \epsilon$. Hence, we can further obtain

$$f(A_1) = \max_{x\in\Delta}\min_{y\in\Delta} x^\mathsf{T} A_1 y \geq \max_{x\in\Delta}\min_{y\in\Delta} x^\mathsf{T} A_2 y - \epsilon = f(A_2) - \epsilon. \tag{E.1}$$

Symmetrically, we can obtain $f(A_2) \geq f(A_1) - \epsilon$. Therefore, we conclude the proof. □

Lemma E.3 ($\epsilon-$Covering)**.** *For any $\epsilon > 0$, the $\epsilon-$covering number $\mathcal{N}_{h,\epsilon}$ of $\overline{\mathcal{Q}}_h$ (and $\underline{\mathcal{Q}}_h$) with respect to ℓ_∞ norm satisfies*

$$\mathcal{N}_{h,\epsilon} \leqslant \left(1 + \frac{4H\sqrt{dK}}{\epsilon}\right)^d \left(1 + \frac{8\beta^2\sqrt{d}}{\epsilon^2}\right)^{d^2},$$

Here the function classes $\overline{\mathcal{Q}}_h$ and $\underline{\mathcal{Q}}_h$ are defined in (A.9)*.*

Proof of Lemma E.3. We only estimate the covering number of $\overline{\mathcal{Q}}_h$. Suppose Q_1 with parameters (w_1, A_1) and Q_2 with parameters (w_2, A_2) are in the function class $\overline{\mathcal{Q}}_h$, then

$$\begin{aligned} \|Q_1 - Q_2\|_\infty &= \sup_{\phi:\|\phi\|\leqslant 1} \left|\Pi_{H-h+1}\left(\phi^\top w_1 + \beta\sqrt{\phi^\top A_1\phi}\right) - \Pi_{H-h+1}\left(\phi^\top w_2 + \beta\sqrt{\phi^\top A_2\phi}\right)\right| \\ &\leqslant \sup_{\phi:\|\phi\|\leqslant 1} \left|\left(\phi^\top w_1 + \beta\sqrt{\phi^\top A_1\phi}\right) - \left(\phi^\top w_2 + \beta\sqrt{\phi^\top A_2\phi}\right)\right| \\ &\leqslant \sup_{\phi:\|\phi\|\leqslant 1} \left|\phi^\top(w_1 - w_2)\right| + \sup_{\phi:\|\phi\|\leqslant 1} \beta\sqrt{|\phi^\top(A_1 - A_2)\phi|} \\ &\leqslant \|w_1 - w_2\| + \beta\sqrt{\|A_1 - A_2\|_F}, \end{aligned}$$

where the second inequality follows from the inequality $\sqrt{x}-\sqrt{y}\leqslant\sqrt{|x-y|}$. Thus $\epsilon/2$-covering of $C_w=\{w\in\mathbb{R}^d:\|w\|\leqslant H\sqrt{dK}\}$ and $\frac{\epsilon^2}{4\beta^2}-$covering of $C_A=\{A\in\mathbb{R}^{d\times d}:\|A\|_F\leqslant\sqrt{d}\}$ are sufficient to form an ϵ-cover of $\overline{\mathcal{Q}}_h$. We obtain the covering number of $\overline{\mathcal{Q}}_h$ satisfies

$$\mathcal{N}_{h,\epsilon}\leqslant\left(1+\frac{4H\sqrt{dK}}{\epsilon}\right)^d\left(1+\frac{8\beta^2\sqrt{d}}{\epsilon^2}\right)^{d^2}.$$

The inequality follows from the standard bound of the covering number of Euclidean Balls (cf. Lemma 2 in Vershynin (2010)). □

Lemma E.4 (Concentration for Self-normalized Processes (Abbasi-Yadkori et al., 2011))**.** *Suppose $\{\epsilon_t\}_{t\geqslant1}$ is a scalar stochastic process generating the filtration $\{\mathcal{F}_t\}_{t\geqslant1}$, and $\epsilon_t|\mathcal{F}_{t-1}$ is zero mean and σ-subGaussian. Let $\{\phi_t\}_{t\geqslant1}$ be an $\mathbb{R}^d$-valued stochastic process with $\phi_t\in\mathcal{F}_{t-1}$. Suppose $\Lambda_0\in\mathbb{R}^{d\times d}$ is positive definite, and $\Lambda_t=\Lambda_0+\sum_{s=1}^t\phi_s\phi_s^\top$. Then for each $\delta\in(0,1)$, with probability at least $1-\delta$, we have*

$$\|\sum_{s=1}^t\phi_s\epsilon_s\|^2_{\Lambda_t^{-1}}\leqslant2\sigma^2log\left(\frac{det(\Lambda_t)^{\frac{1}{2}}}{\delta det(\Lambda_0)^{\frac{1}{2}}}\right),\quad\forall t\geqslant0.$$

Lemma E.5 (Matix Bernstein Inequality)**.** *Supposed that $\{A_k\}_{k=1}^n$ are independent random matrix in $\mathbb{R}^{d_1\times d_2}$. They satisfy $\mathbb{E}[A_k]=0$ and $\|A_k\|_{op}\leqslant L$. Let $Z=\sum_{k=1}^n A_k$ and*

$$v(Z)=\max\left\{\left\|\mathbb{E}[ZZ^\top]\right\|_{op},\left\|\mathbb{E}[Z^\top Z]\right\|_{op}\right\}=\max\left\{\left\|\sum_{k=1}^n\mathbb{E}[A_kA_k^\top]\right\|_{op},\left\|\sum_{k=1}^n\mathbb{E}[A_k^\top A_k]\right\|_{op}\right\}.$$

We have,

$$\mathbb{P}(\|Z\|_{op}\geqslant t)\leqslant(d_1+d_2)\cdot\exp\left(-\frac{t^2/2}{v(Z)+L/3\cdot t}\right),\quad\forall t>0.$$

Proof. See Theorem 1.6.2 of Tropp (2015) for detailed proof. □