WHEN ARE OFFLINE TWO-PLAYER ZERO-SUM MARKOV GAMES SOLVABLE?

Qiwen Cui
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
qwcui@cs.washington.edu

Simon S. Du
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
ssdu@cs.washington.edu

ABSTRACT

We study what dataset assumption permits solving offline two-player zero-sum Markov games. In stark contrast to the offline single-agent Markov decision process, we show that the single strategy concentration assumption is insufficient for learning the Nash equilibrium (NE) strategy in offline two-player zero-sum Markov games. On the other hand, we propose a new assumption named unilateral concentration and design a pessimism-type algorithm that is provably efficient under this assumption. In addition, we show that the unilateral concentration assumption is necessary for learning an NE strategy. Furthermore, our algorithm can achieve minimax sample complexity without any modification for two widely studied settings: dataset with uniform concentration assumption and turn-based Markov games. Our work serves as an important initial step towards understanding offline multi-agent reinforcement learning.

1 INTRODUCTION

Promising empirical advances have been achieved in reinforcement learning (RL), including mastering the games of Go (Silver et al., 2016), Poker (Brown et al., 2017), real-time strategy games (Vinyals et al., 2019) and robotic control (Kober et al., 2013). Notably, many of these successes lie in the domain of multi-agent reinforcement learning (MARL). MARL is about multiple agents interacting in a shared environment, and each of them aims to maximize its own long-term reward. During the learning process, each agent not only needs to identify the environment dynamic but also needs to compete/cooperate with other agents. One important subarea of MARL is offline MARL. In many practical scenarios, we only have access to the offline data or it is too expensive to frequently change the policy (Zhang et al., 2021a). While there are plenty of empirical works on offline MARL (Pan et al., 2021; Jiang & Lu, 2021), the theoretical understanding about offline MARL is still very limited. In this work, we take an initial step towards understanding when offline MARL is provably solvable.

We consider the two-player zero-sum Markov games, where two players sequentially select actions in a Markovian environment and the first player aims to maximize the total reward while the second player aims to minimize it. In the offline setting, we have access to a fixed dataset collected by a (possibly unknown) exploration policy and the target is to find a (near-)Nash equilibrium (NE) strategy of the underlying two-agent zero-sum Markov games.

One of the main difficulties in offline RL is distribution shift, i.e. the dataset distribution is different from the distribution induced by the optimal policy. It is important to understand what is the minimal dataset distribution assumption that permits offline RL. For single-agent offline RL, it is shown that the pessimism principle allows policy optimization with *single policy concentration*, i.e. the dataset only covers the optimal policy (Jin et al., 2021b; Zanette et al., 2021; Yin & Wang, 2021; Rashidinejad et al., 2021). This assumption is necessary, as it is impossible to learn the optimal policy if it is not covered by the dataset. However, the dataset coverage assumption for MARL is far from clear. In this work, we want to answer the following question:

What is the minimal dataset coverage assumption that permits learning an NE strategy in offline two-player zero-sum Markov games?

Generally speaking, MARL is much more difficult than single-agent RL due to the following two reasons. First, MARL is known to suffer from the *non-stationary* property, i.e. agents will affect the others during the learning process (Zhang et al., 2021a). Specifically, the performance may decline if each agent simultaneously tries to improve their own policy depends on others' current policies. In addition, multiple agents incur complicated statistical dependence that makes the theoretical analysis difficult. A line of works studies Markov games with online sampling oracle (Bai et al., 2020; Bai & Jin, 2020; Liu et al., 2021) or generative model oracle (Sidford et al., 2020; Zhang et al., 2020; Cui & Yang, 2020), where specialized techniques are developed to tackle the above difficulties. In this paper, we give the first analysis on Markov games in the offline setting.

1.1 Main Contributions

First, as a warm up, we consider the most natural extension of the single policy concentration assumption in single-agent RL: the dataset covers the NE strategy in offline two-player zero-sum games. However, in Section 3.1, we prove that it is *impossible to learn the NE strategy under such assumption.* We construct a pair of hard Markov games such that there exists a dataset that covers the NE strategies in both Markov games but no algorithm can distinguish between these two Markov games and their NE strategies. This hardness result refutes the conjecture that single strategy concentration works for Markov games.

Second, we propose an assumption named *unilateral concentration*, which posits that for all strategy μ, ν, strategy pairs (μ^*, ν) and (μ, ν^*) are covered by the dataset, where μ is the strategy for the first (max) player, ν is the strategy for the second (min) player, and (μ^*, ν^*) is the NE strategy. In Section 3.2, we prove that NE strategy is not learnable even if this assumption is only slightly violated. The intuition behind these hardness results is that to identify an NE strategy, the algorithm has to compare it with the strategy that one player uses any other strategies as a reference.

Third, we provide positive results showing that NE strategy is PAC learnable under the unilateral concentration assumption. Combined with the hardness results above, we can conclude that unilateral concentration assumption is the *necessary and sufficient dataset coverage assumption for solving offline Markov games*. Our algorithm is based on the pessimism principle that we maintain both a pessimistic and optimistic estimate for each player, respectively. We show that our algorithm can achieve $\widetilde{O}(\sqrt{C^*SABH^3/n})$ performance gap under unilateral concentration assumption, where C^* quantifies the coverage of the dataset, S is the number of states, A is the number of max player's action, B is the number of min player's action, H is the horizon and n is the number of samples.

Fourth, we show that our algorithm is *minimax optimal* when the dataset satisfies a stronger assumption, uniform concentration, or the Markov game is turn-based. These are two widely studied settings in the RL community. Uniform concentration assumes that all state-action pairs is covered by the dataset and turn-based Markov game is a variant of zero-sum Markov game where two players select actions in turns instead of simultaneously. Although uniform concentration is about the dataset structure and turn-based Markov game is about the environment structure, our algorithm can adapt to both of them without any modification and achieves minimax sample complexity.

Our algorithm is motivated by the Bernstein-type bonus and reference advantage function techniques in Xie et al. (2021) while we make novel adaptations, monotonic update and self-bounding technique, to realize them in Markov games. Monotonic update allows a sandwich-type argument that bounds the reference function and further bounds the variance term. Self-bounding technique is utilized to bound the performance gap by itself and then solve the inequality to derive the final bound on performance gap.

To summarize, (1) we identify the minimal dataset coverage assumption that allows learning the NE strategy in Markov games; (2) we propose a pessimism-based algorithm that achieves polynomial sample complexity based on novel Markov game techniques; and (3) we further show the algorithm is minimax optimal under the uniform concentration assumption or in turn-based Markov games.

2 Preliminaries

2.1 Two-Player Zero-sum Markov Game

Zero-sum Markov game (MG) generalizes single-agent MDP to two-agent case where one agent aims to maximize the total reward while the other one aims to minimize it. A finite-horizon time-inhomogeneous zero-sum Markov game is described by the tuple $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space of the first (max) player, $\mathcal{B}$ is the action space of the second (min) player, $P = (P_1, P_2, \cdots, P_H), P_h \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{B}| \times |\mathcal{S}|}, \forall h \in [H]$ is the (unknown) transition probability matrix for time step h, $r = (r_1, r_2, \cdots, r_H), r_h \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{B}|}, \forall h \in [H]$ is the (unknown) deterministic reward vector and H is the horizon length. [*] At each timestep h and state s_h, if the max player chooses an action a_h and the min player chooses an action b_h, then the next state at timestep $h+1$ follows the distribution $s_{h+1} \sim P_h(\cdot|s_h, a_h, b_h)$ and both players receive a reward $r_h(s_h, a_h, b_h)$. Both players sequentially choose H actions and at each timestep, the action is chosen *simultaneously* and then it is revealed to both players. We assume that we have a fixed initial state s_1 and it is straightforward to generalize our result to the case where the initial state is sampled from a fixed distribution.[†]

Turn-based Markov game is an important subclass of (simultaneous-move) Markov game, where the max player takes the action first and the min player can take the action after observing the opponent's action. It is a widely studied setting (Sidford et al., 2020; Cui & Yang, 2020; Bai & Jin, 2020) and we will provide minimax sample complexity result for this setting in Section 4.3.

We denote a strategy pair as $\pi = (\mu, \nu)$, where $\mu = (\mu_1, \mu_2, \cdots, \mu_H), \mu_h : \mathcal{S} \to \Delta^{\mathcal{A}}, \forall h \in [H]$ is the strategy of the first player and $\nu = (\nu_1, \nu_2, \cdots, \nu_H), \nu_h : \mathcal{S} \to \Delta^{\mathcal{B}}, \forall h \in [H]$ is the strategy of the second player, where $\Delta^{\mathcal{X}}$ is the probability simplex on the finite set $\mathcal{X}$. A deterministic strategy is a strategy that maps state to a single point distribution. We define the state value function and state-action value function for a strategy pair π similarly as in single-agent MDP:

$$V_h^\pi(s_h) := \mathbb{E}\left[\sum_{t=h}^{H} r(s_t, a_t, b_t)|\pi, s_h\right],$$

$$Q_h^\pi(s_h, a_h, b_h) := \mathbb{E}\left[\sum_{t=h}^{H} r(s_t, a_t, b_t)|\pi, s_h, a_h, b_h\right].$$

If the second player's strategy ν is fixed, then the MG degenerates to an MDP and we call the optimal policy in this MDP as the best response strategy $\mathrm{br}_1(\nu)$. Similarly we can define the $\mathrm{br}_2(\mu)$ as the best response for the second player. We will ignore the subscript in br_1 and br_2 when it is clear in the context. For all $h \in [H], s_h \in \mathcal{S}$, we define

$$V_h^{*,\nu}(s_h) := V_h^{\mathrm{br}(\nu),\nu}(s_h) = \max_\mu V_h^{\mu,\nu}(s_h),$$

$$V_h^{\mu,*}(s_h) := V_h^{\mu,\mathrm{br}(\mu)}(s_h) = \min_\nu V_h^{\mu,\nu}(s_h).$$

It is well known that Nash equilibrium (NE) strategy $\pi^* = (\mu^*, \nu^*)$, i.e. a strategy pair such that no player can benefit from switching its own strategy, exists for zero-sum Markov games with a unique value function (Shapley, 1953). In other words, μ^* and ν^* are the best responses to each other. We define $V_h^* := V_h^{\mu^*,\nu^*}$ for all $h \in [H]$. The following weak duality property holds for all strategy pair (μ, ν) in MG:

$$V_h^{\mu,*} \le V_h^* \le V_h^{*,\nu}, \forall h \in [H].$$

For a strategy pair $\pi = (\mu, \nu)$, we can then define the corresponding duality gap as

$$\mathrm{Gap}(\pi) = V_1^{*,\nu}(s_1) - V_1^{\mu,*}(s_1).$$

The duality gap is always non-negative and the NE strategy has zero duality gap $\mathrm{Gap}(\pi^*) = 0$. Duality gap measures how well a strategy pair approximates the NE. We say a strategy pair π is an ϵ-approximate NE if $\mathrm{Gap}(\pi) \le \epsilon$.

[*] While we assume deterministic rewards for simplicity, our results can be straightforwardly generalized to unknown stochastic rewards, as the major difficulty is in learning the transitions rather than learning the rewards.

[†] Stochastic initial state is equivalent to an MDP with deterministic initial state by creating a dummy initial state which transit to the next state following that initial state distribution.

2.2 Offline Two-Player Zero-Sum Game

In offline RL, we are given an offline dataset $D = \{(s_h^\tau, a_h^\tau, b_h^\tau, r_h^\tau, s_{h+1}^\tau)\}_{\tau\in[n]}^{h\in[H]}$ and we cannot do any further sampling (Kakade, 2003). We assume that the dataset is sampled from some exploration policy $\rho = (\rho_1, \rho_2, \cdots, \rho_H), \rho_h : \mathcal{S} \to \Delta^{\mathcal{A}\times\mathcal{B}}, \forall h \in [H]$.[‡] The target of offline MG is to find an approximate NE with a small duality gap by utilizing the given dataset D. We use $d_h^\pi(s, a, b)$ to denote the probability of s, a, b appears at timestep h in the trajectory generated by strategy π for all $h \in [H]$. The dataset distribution $d_h^\rho(s, a, b)$ is defined similarly. A state-action pair (s, a, b) at timestep h is covered by strategy π if and only if $d_h^\pi(s, a, b) > 0$. Strategy π is covered by dataset generated by exploration strategy ρ if and only if for all (s, a, b) covered by π, it is covered by ρ. In other words, we have

$$\frac{d_h^\pi(s, a, b)}{d_h^\rho(s, a, b)} < \infty, \forall h \in [H], (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}. \tag{1}$$

The sample complexity guarantee will depend on this ratio.

Dataset Coverage Assumptions Below we list three different dataset coverage assumptions for Markov games.

Assumption 2.1. (Single strategy concentration) The NE strategy (μ^*, ν^*) is covered by the dataset.

Assumption 2.2. (Unilateral concentration) For all strategy μ and ν, (μ, ν^*) and (μ^*, ν) are covered by the dataset, where (μ^*, ν^*) is the NE strategy.

Assumption 2.3. (Uniform concentration) For all $h \in [H]$ and $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, (s, a, b) at timestep h is covered by the dataset.

Assumption 2.1 is the weakest assumption and is the most straightforward extension of the single policy concentration in single-agent RL (Rashidinejad et al., 2021). Assumption 2.3 generalizes the uniform policy concentration in single-agent RL (Yin et al., 2020). Assumption 2.2 is sandwiched by Assumption 2.1 and Assumption 2.3 as Assumption 2.2 implies Assumption 2.1 and Assumption 2.3 implies Assumption 2.2. In this work, we will show that Assumption 2.2 is the minimal dataset coverage assumption that allows NE learning and we provide sample complexity bounds that depends on the density ratio (1).[§]

Notations. We use $\mathrm{Var}_{P(s,a,b)}(V)$ to denote the variance of the random variable $V(s')$ where $s' \sim P(\cdot|s, a, b)$ and $\mathrm{Var}_P(V) \in \mathbb{R}^{SAB}$ to denote a vector whose (s, a, b) component is $\mathrm{Var}_{P(s,a,b)}(V)$. We define $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. In addition, if a is a vector and b is a scalar, the operation is taken on each element of a: $[a \vee b]_i = a_i \vee b$. For two vector $a \in \mathbb{R}^n$, $b \in \mathbb{R}^n$, we use $\frac{a}{b} \in \mathbb{R}^n$ to denote the element-wise division: $\left[\frac{a}{b}\right]_i = \frac{a_i}{b_i}$. In addition, if a is scalar, we still use $\frac{a}{b} \in \mathbb{R}^n$ to denote the element-wise division: $\left[\frac{a}{b}\right]_i = \frac{a}{b_i}$.

3 Impossibility Results

In this section, we will provide two hard instances. To begin with, we show that single strategy concentration is not enough for NE strategy learning. In addition, we show that no assumption weaker than unilateral concentration allows NE strategy learning.

3.1 Warm Up: Single Strategy Coverage is Insufficient for NE Identification

To illustrate the hardness of offline MARL, below we construct an example showing that single strategy concentration is insufficient for NE identification. We consider bandits game, i.e. Markov game with horizon $H = 1$. The result can be easily generalized to arbitrary horizon by setting the

[‡] For simplicity we assume the exploration policy is Markovian. It is actually unnecessary because our algorithm and analysis only depend on the distribution of the dataset instead of this Markovian property. See Jin et al. (2021b) for details.

[§] Note that there could be different minimal assumption as the assumption set is a partially ordered set. Here 'minimal' means Assumption 2.2 allows NE learning while no weaker assumption allows doing so.

reward to be 0 in horizons other than $h = 1$. We consider a class of bandit game and dataset such that the NE strategy is well covered and we show that no algorithm can identify the NE strategy for all bandits game and dataset in this class.

Theorem 3.1. *Define a class $\mathcal{X}$ of bandits game M and exploration strategy ρ that consists of all M and ρ pairs which satisfy for all $s \in \mathcal{S}$ and $a, b \in \operatorname{support}(\pi^*(\cdot, \cdot|s))$,*

$$\frac{d^{\pi^*}(s, a, b)}{d^{\rho}(s, a, b)} \le 2,$$

*where π^* is the NE strategy in M. For any algorithm **ALG**, there exists $(M, \rho) \in \mathcal{X}$ such that the output of the algorithm **ALG** is at most a $1/2$-approximate NE strategy no matter how many data are collected.*

Theorem 3.1 denies the possibility that single strategy concentration allows NE strategy learning. The reason that single policy concentration works in single-agent RL instead of MARL is that we can use the data of two actions to decide which one is better in MDP but we cannot use the data of two action pairs to decide which pair is closer to the NE strategy in Markov games as NE identification needs other action pairs to be the references.

3.2 No Assumption Weaker Than Assumption 2.2 Is Sufficient for NE Learning

In this section, we will show that it is impossible to learn the NE strategy even if Assumption 2.2 is slightly violated. To begin with, we consider the following deterministic unilateral concentration assumption.

Assumption 3.2. (Deterministic unilateral concentration) For all deterministic strategy μ and ν, (μ, ν^*) and (μ^*, ν) are covered by the dataset, where (μ^*, ν^*) is the NE strategy.

Immediately we can tell that Assumption 3.2 is satisfied under Assumption 2.2. These two assumptions are equivalent, which is shown by Proposition 3.3. The intuition is that any stochastic strategy can be viewed as a combination of several deterministic strategies so it is covered by the dataset if all the deterministic strategies are covered.

Proposition 3.3. *If for all deterministic strategy μ and ν, (μ, ν^*) and (μ^*, ν) are covered by the dataset, then we have for all (possibly stochastic) strategy μ' and ν', (μ', ν^*) and (μ^*, ν') are covered by the dataset.*

Theorem 3.4. *Define a class $\mathcal{X}$ of bandits game M and exploration strategy ρ that consists of all M and ρ pairs satisfying that there exists at most one deterministic strategy μ or one deterministic strategy ν such that (μ, ν^*) or (μ^*, ν) is not covered and for all other deterministic strategy μ', ν', the density ratio is bounded*

$$\frac{d_h^{\mu^*, \nu'}(s, a, b)}{d_h^{\rho}(s, a, b)} \le 2|\mathcal{A}| + 2|\mathcal{B}|, \frac{d_h^{\mu', \nu^*}(s, a, b)}{d_h^{\rho}(s, a, b)} \le 2|\mathcal{A}| + 2|\mathcal{B}|,$$

*for all $h \in [H]$. For any algorithm **ALG**, there exists $(M, \rho) \in \mathcal{X}$ such that the output of the algorithm **ALG** is at most a 0.25-approximate NE strategy no matter how many data is collected.*

Remark 3.5. We can easily adapt this instance to arbitrary action space by setting all the other rewards to be 0.5 and the exploration strategy ρ to be the uniform distribution on (a_i, b_j) such that $(i, j) \in \{(i, j) : i \in \{1, 2\} \text{ or } j \in \{1, 2\}, (i, j) \ne (2, 1)\}$.

Remark 3.6. It is straightforward to verify that the hard instance in Theorem 3.4 also holds for turn-based Markov games. As a result, no assumption weaker than Assumption 2.2 is sufficient for NE learning in turn-based Markov games.

Theorem 3.4 suggests that no assumption weaker than Assumption 3.2 allows NE strategy learning. As Assumption 2.2 and Assumption 3.2 are equivalent, no assumption weaker than Assumption 2.2 allows NE strategy learning.

4 Provably Efficient Algorithm under Unilateral Concentration

In this section, we show that it is indeed possible to learn the NE with the unilateral concentration assumption. We propose a novel algorithm called Pessimistic and Optimistic Value Iteration (POVI), which adapts the pessimism principle in single-agent RL to Markov games. Our sample complexity result depends on the following quantity name *unilateral concentrability*:

Definition 4.1. (Unilateral concentrability)

$$C^* := \max_{h,(s,a,b),\mu,\nu} \left\{ \frac{d_h^{\mu^*,\nu}(s,a,b)}{d_h^{\rho}(s,a,b)}, \frac{d_h^{\mu,\nu^*}(s,a,b)}{d_h^{\rho}(s,a,b)} \right\}.$$

By definition, C^* is finite if Assumption 2.2 is satisfied. Note that there could be multiple different NE strategies, which correspond to different C^*. Our guarantee holds all of them without the knowledge of NE or C^*.

4.1 Hoeffding-type Algorithm with Data Splitting

To illustrate our main algorithm design ideas, we first propose an algorithm with Hoeffding-type bonus and random data splitting. Given a dataset $\mathcal{D} = \left\{(s_h^k, a_h^k, b_h^k, r_h^k, s_{h+1}^k)\right\}_{k,h=1}^{n,H}$, we denote $n_h(s,a,b) = \sum_{k=1}^n \mathbf{1}\left((s_h^k, a_h^k, b_h^k) = (s,a,b)\right)$ to be the number of times that (s,a,b) is visited at timestep h. We set the empirical reward as

$$\widehat{r}_h(s,a,b) = r_h(s,a,b), \tag{2}$$

and the empirical transition kernel as

$$\widehat{P}_h(s'|s,a,b) = \frac{\sum_{k=1}^n \mathbf{1}\left((s_h^k, a_h^k, b_h^k, s_{h+1}^k) = (s,a,b,s')\right)}{\sum_{k=1}^n \mathbf{1}\left((s_h^k, a_h^k, b_h^k) = (s,a,b)\right)}, \tag{3}$$

if $n_h(s,a,b) \geq 1$ and $\widehat{r}_h(s,a,b) = 0$, $\widehat{P}_h(s'|s,a,b) = 1/S$ otherwise. In addition, we use $n_h \in \mathbb{R}^{SAB}$ to denote a vector such that $[n_h]_{s,a,b} = n_h(s,a,b)$.

Now we explain Algorithm 1 in details. First, we split the dataset $\mathcal{D}$ into H small datasets $\{\mathcal{D}_h\}_{h=1}^H$ with the same size. Then we use $\mathcal{D}_h$ to estimate the reward and the transition matrix at timestep h. The data splitting scheme is to remove the dependence between each timestep. Then the value function is estimated via a value-iteration-type algorithm. At each timestep, we maintain both optimistic and pessimistic estimates by adding/minusing a Hoeffding-type bonus. We use the following Hoeffding-type bonus:

$$\underline{b}_h(s_h,a_h,b_h) = \overline{b}_h(s_h,a_h,b_h) = 4\sqrt{\frac{H^2\iota}{n_h(s,a,b)\vee 1}}, \tag{4}$$

where $\iota = \log(HSAB/\delta)$. Then we can compute the pessimistic estimate $\overline{Q}$ and optimistic estimate $\underline{Q}$:

$$\underline{Q}_h = \left(\widehat{r}_h + (\widehat{P}_h \cdot \underline{V}_{h+1}) - \underline{b}_h\right) \vee 0, \tag{5}$$

$$\overline{Q}_h = \left(\widehat{r}_h + (\widehat{P}_h \cdot \overline{V}_{h+1}) + \overline{b}_h\right) \wedge (H-h+1). \tag{6}$$

The pessimistic estimate is for the max player, which mimics the pessimism in single-agent RL. The optimistic estimate is for the min player, which is also a kind of pessimism as the min player's target is to minimize the reward. We compute the NE strategy of the matrix game $\underline{Q}(s,\cdot,\cdot)$ and $\overline{Q}(s,\cdot,\cdot)$ respectively and use the NE value to be the state value $\underline{V}(s)$ and $\overline{V}(s)$. Note that we only solve a zero-sum matrix game, which is computationally efficient (Chen & Deng, 2006).

Theorem 4.2. *Suppose Assumption 2.2 holds. For any $0 < \delta < 1$ and policy μ, ν, with probability $1-\delta$, the pessimistic value $\underline{V}_h$ and optimistic value $\overline{V}_h$ of Algorithm 1 satisfies the following arguments for all $h \in [H]$*

$$\mathbb{E}_{\mu^*,\nu}\left[V_h^*(s_h) - \underline{V}_h(s_h)\right] \leq \widetilde{O}\left(\sqrt{C^*SABH^5/n}\right),$$

$$\mathbb{E}_{\mu,\nu^*}\left[\overline{V}_h(s_h) - V_h^*(s_h)\right] \leq \widetilde{O}\left(\sqrt{C^*SABH^5/n}\right),$$

where s_h is sampled from the trajectory following the strategy in the expectation.

Theorem 4.2 provides polynomial bounds on the error of the value estimates in Algorithm 1. It can directly imply the following performance gap bound. In addition, it provides guarantees for the reference function that will be utilized in the next section.

Corollary 4.3. *Suppose Assumption 2.2 holds. For any $0 < \delta < 1$ and $n \geq \widetilde{O}(\sqrt{C^* SABH^5})$, with probability $1 - \delta$, the output policy $\pi = (\underline{\mu}, \overline{\nu})$ of Algorithm 1 satisfies*

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \leq \widetilde{O}\left(\sqrt{C^* SABH^5/n}\right).$$

Theorem 4.3 shows that the output strategy of Algorithm 1 is an $\widetilde{O}\left(\sqrt{C^* SABH^5/n}\right)$-approximate NE. The parameter C^* measures how the exploration strategy ρ covers the strategy space (μ^*, ν) and (μ, ν^*) for all μ and ν.

4.2 Bernstein-type Algorithm with Reference Advantage Function Decomposition

In this section, we will derive an improved performance gap bound $\widetilde{O}\left(\sqrt{C^* SABH^3/n}\right)$. The extra H^2 is shaved by using Bernstein-type bonus and reference advantage decomposition technique, which is motivated from Xie et al. (2021). However, we want to emphasize that zero-sum Markov game is substantially different from MDP and requires novel adaptation, which we will describe later.

Algorithm 2 is different from Algorithm 1 in two aspects. First, we use the reference advantage decomposition to remove an H factor. The dataset is split into three subset with equal size $\mathcal{D}_{\mathrm{ref}}$, $\mathcal{D}_0$, $\mathcal{D}_1$, and $\mathcal{D}_1$ is further split into H subset with equal size $\{\mathcal{D}_{h,1}\}_{h=1}^H$. We run algorithm 1 on dataset $\mathcal{D}_{\mathrm{ref}}$ and we can obtain pessimistic value estimate $\underline{V}_{\mathrm{ref}}$ and optimistic value estimate $\overline{V}_{\mathrm{ref}}$ with guarantees by Theorem 4.2. Then we use dataset $\mathcal{D}_0$ to estimate $P_h \underline{V}_{h+1}^{\mathrm{ref}}$ and dataset $\mathcal{D}_{h,1}$ to estimate $P_h(\underline{V}_{h+1} - \underline{V}_{h+1}^{\mathrm{ref}})$. Second, we use a Bernstein-type bonus to remove another H factor. The our updating formulas of $\underline{Q}_h$ and $\overline{Q}_h$ are

$$\underline{Q}_h = \underline{Q}_h^{\mathrm{ref}} \vee [\widehat{r}_{h,0} + (\widehat{P}_{h,0} \cdot \underline{V}_{h+1}^{\mathrm{ref}}) - \underline{b}_{h,0} + (\widehat{P}_{h,1} \cdot (\underline{V}_{h+1} - \underline{V}_{h+1}^{\mathrm{ref}})) - \underline{b}_{h,1}], \tag{7}$$

$$\overline{Q}_h = \overline{Q}_h^{\mathrm{ref}} \wedge [\widehat{r}_{h,0} + (\widehat{P}_{h,0} \cdot \underline{V}_{h+1}^{\mathrm{ref}}) - \overline{b}_{h,0} + (\widehat{P}_{h,1} \cdot (\overline{V}_{h+1} + \overline{V}_{h+1}^{\mathrm{ref}})) + \overline{b}_{h,1}], \tag{8}$$

where we truncate by the reference function to ensure monotonic update so that $\underline{Q}_h$ and $\overline{Q}_h$ are more accurate pessimistic/optimistic estimate compared with the reference function $\underline{Q}_h$ and $\overline{Q}_h$. The bonus functions are defined as

$$\begin{aligned} \underline{b}_{h,0} &= c\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,0}}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,0} \vee 1}} + \frac{H\iota}{n_{h,0} \vee 1}\right), \\ \overline{b}_{h,0} &= c\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,0}}(\overline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,0} \vee 1}} + \frac{H\iota}{n_{h,0} \vee 1}\right), \end{aligned} \tag{9}$$

$$\begin{aligned} \underline{b}_{h,1} &= c\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,1}}(\underline{V}_{h+1} - \underline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,1} \vee 1}} + \frac{H\iota}{n_{h,1} \vee 1}\right), \\ \overline{b}_{h,1} &= c\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,1}}(\overline{V}_{h+1} - \overline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,1} \vee 1}} + \frac{H\iota}{n_{h,1} \vee 1}\right), \end{aligned} \tag{10}$$

where c is some universal constant and $\mathrm{Var}_{\widehat{P}_{h,0}}(V)$, $\mathrm{Var}_{\widehat{P}_{h,1}}(V)$, $n_{h,0}$, $n_{h,1}$ are all SAB-dimension vectors and the operations are element-wise.

Theorem 4.4. *Suppose Assumption 2.2 holds. For any $0 < \delta < 1$ and $n \geq \widetilde{O}(\sqrt{C^* SABH^4})$, with probability $1-\delta$, the output policy $\pi = (\underline{\mu}, \overline{\nu})$ of Algorithm 2 satisfies*

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \leq \widetilde{O}\left(\sqrt{C^* SABH^3/n}\right).$$

As MDP is degenerated Markov game with one player having a fixed action, Markov game inherits the lower bounds of MDP. Comparing with the lower bound $\widetilde{\Omega}\left(\sqrt{C^* SH^3/n}\right)$ (Xie et al., 2021), our bound is already tight in C^*, S, H. The extra AB factor is from the Cauchy-Schwarz inequality and the fact that the NE of zero-sum Markov game can be mixed strategy. This makes the bound different from MDP where deterministic optimal policy always exists. It is unknown whether the AB factor is removable and we leave it to future work.

4.3 Minimax Optimal Sample Complexity Bounds

In this section, we show that Algorithm 2 directly adapts to two popular settings, i.e. Assumption 2.3 (uniform concentration assumption) and turn-based Markov game. In addition, minimax sample complexity can be achieved under both settings. The proof is deferred to Appendix F.

Theorem 4.5. *Suppose Assumption 2.3 holds and $d_m = \min\{d_h^\rho(s,a,b) : h \in [H], (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}\}$. For any $0 < \delta < 1$ and $n \geq \widetilde{O}\left(\sqrt{H^4/d_m}\right)$, with probability $1-\delta$, the output policy $\pi = (\underline{\mu}, \overline{\nu})$ of Algorithm 2 satisfies*

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \leq \widetilde{O}\left(\sqrt{H^3/(nd_m)}\right).$$

This bound has no explicit dependence on AB because the Cauchy-Schwarz inequality can be applied on $d_h^{\mu^*,\underline{\nu}}$ instead of $\sqrt{d_h^{\mu^*,\underline{\nu}}}$ (See the proof of Theorem F.1). As the lower bound $\widetilde{\Omega}\left(\sqrt{H^3/(nd_m)}\right)$ for MDP (Yin & Wang, 2021) is the lower bound for Markov games, Algorithm 2 achieves minimax sample complexity under assumption 2.3.

Theorem 4.6. *Suppose Assumption 2.2 holds for a turn-based Markov game. For any $0 < \delta < 1$ and $n \geq \widetilde{O}(\sqrt{C^* SH^4})$, with probability $1-\delta$, the output policy $\pi = (\underline{\mu}, \overline{\nu})$ of Algorithm 2 satisfies*

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \leq \widetilde{O}\left(\sqrt{C^* SH^3/n}\right).$$

As the lower bound is $\widetilde{\Omega}\left(\sqrt{C^* SH^3/n}\right)$ (Xie et al., 2021), Algorithm 2 can achieve the minimax sample complexity for turn-based Markov games under assumption 2.2. The difference between turn-based Markov game and simultaneous-move Markov game is that the former one has pure NE strategy, which saves the AB factor when Cauchy-Schwarz inequality is applied (See the proof of Theorem F.6).

5 Conclusion

In this work, we study the minimal dataset coverage assumption for NE learning in two-player zero-sum Markov games. We show that single strategy concentration is not enough for NE learning. Instead, we find a minimal coverage assumption for NE learning and design an algorithm with sample complexity tight in $C^*, \mathcal{S}, H$ under such assumption based on novel techniques. In addition, the algorithm can achieve minimax sample complexity in certain settings. We believe this work can shed new light on offline MARL.

Here we list several open problems for future work. One direction is to find the minimax sample complexity of offline Markov games under unilateral concentration. Importantly, it is unclear whether AB factor can be reduced to $A + B$ as in the online setting Bai et al. (2020). Another direction is to design decentralized algorithm for offline MARL. The answer to this question is especially crucial if we want to further study offline MARL with a large number of agents and we do not want the sample complexity scales exponentially with the number of agents. Lastly, in this paper we only focus on the most fundamental tabular setting. It is natural to ask how to generalize our findings, especially the unilateral concentration assumption, to the function approximation setting.

REFERENCES

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

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

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

Noam Brown, Tuomas Sandholm, and Strategic Machine. Libratus: The Superhuman AI for No-Limit Poker. In *IJCAI*, pp. 5226–5228, 2017.

Xi Chen and Xiaotie Deng. Settling the complexity of two-player nash equilibrium. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 261–272. IEEE, 2006.

Qiwen Cui and Lin F Yang. Minimax sample complexity for turn-based stochastic game. *arXiv preprint arXiv:2011.14267*, 2020.

Zeyu Jia, Lin F Yang, and Mengdi Wang. Feature-based q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Jiechuan Jiang and Zongqing Lu. Offline decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:2108.01832*, 2021.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is Pessimism Provably Efficient for Offline RL? *arXiv:2012.15085 [cs, math, stat]*, May 2021b. URL `http://arxiv.org/abs/2012.15085`. arXiv: 2012.15085.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

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. Publisher: SAGE Publications Sage UK: London, England.

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*, pp. 7001–7010. PMLR, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. *arXiv preprint arXiv:2111.11188*, 2021.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. *arXiv:2103.12021 [cs, math, stat]*, March 2021. URL `http://arxiv.org/abs/2103.12021`. arXiv: 2103.12021.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S. Du, and Sujay Sanghavi. Nearly Horizon-Free Offline Reinforcement Learning. *arXiv:2103.14077 [cs, stat]*, October 2021. URL `http://arxiv.org/abs/2103.14077`. arXiv: 2103.14077.

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

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002. PMLR, 2020.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. ISSN 1476-4687. doi: 10.1038/nature16961. URL `https://doi.org/10.1038/nature16961`.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pp. 880–887, 2005.

Masatoshi Uehara and Wen Sun. Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage. *arXiv:2107.06226 [cs, stat]*, October 2021. URL `http://arxiv.org/abs/2107.06226`. arXiv: 2107.06226.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation Learning for Online and Offline RL in Low-rank MDPs. *arXiv:2110.04652 [cs, stat]*, November 2021. URL `http://arxiv.org/abs/2110.04652`. arXiv: 2110.04652.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, and others. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2, 2019.

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*, pp. 3674–3682. PMLR, 2020a.

Tengyang Xie and Nan Jiang. Batch Value-function Approximation with Only Realizability. *arXiv:2008.04990 [cs, stat]*, June 2021. URL `http://arxiv.org/abs/2008.04990`. arXiv: 2008.04990.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling. *arXiv:1906.03393 [cs, stat]*, March 2020b. URL `http://arxiv.org/abs/1906.03393`. arXiv: 1906.03393.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *arXiv preprint arXiv:2106.04895*, 2021.

Ming Yin and Yu-Xiang Wang. Towards Instance-Optimal Offline Reinforcement Learning with Pessimism. *arXiv:2110.08695 [cs, stat]*, October 2021. URL `http://arxiv.org/abs/2110.08695`. arXiv: 2110.08695.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-Optimal Provable Uniform Convergence in Offline Policy Evaluation for Reinforcement Learning. *arXiv:2007.03760 [cs, stat]*, December 2020. URL `http://arxiv.org/abs/2007.03760`. arXiv: 2007.03760.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-Optimal Offline Reinforcement Learning via Double Variance Reduction. *arXiv:2102.01748 [cs, stat]*, February 2021. URL `http://arxiv.org/abs/2102.01748`. arXiv: 2102.01748.

Andrea Zanette, Martin J. Wainwright, and Emma Brunskill. Provable Benefits of Actor-Critic Methods for Offline Reinforcement Learning. *arXiv:2108.08812 [cs]*, August 2021. URL `http://arxiv.org/abs/2108.08812`. arXiv: 2108.08812.

Kaiqing Zhang, Sham M Kakade, Tamer Başar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

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

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 2021b.

A Algorithms

Algorithm 1 Pessimistic and Optimistic Value Iteration

```
Input: Offline dataset $\mathcal{D} = \big\{(s_h^k, a_h^k, b_h^k, r_h^k, s_{h+1}^k)\big\}_{k,h=1}^{n,H}$. Failure Probability $\delta$.
Initialization: Set $\underline{V}_{H+1}(\cdot) = \overline{V}_{H+1}(\cdot) = 0$. Randomly split the dataset $\mathcal{D}$ into $\{\mathcal{D}_h\}_{h=1}^H$ with $|\mathcal{D}_h| = n/H$. Set $\widehat{P}_h$, $\widehat{r}_h$, $\underline{b}_h$ and $\overline{b}_h$ as (2), (3) and (4) using the dataset $\mathcal{D}_h$ for all $h \in [H]$.
for time $h = H, H-1, \dots, 1$ do
    Set $\underline{Q}_h(\cdot,\cdot,\cdot)$ as (5).
    Compute the NE of $\underline{Q}_h(\cdot,\cdot,\cdot)$ as $(\underline{\mu}_h(\cdot), \underline{\nu}_h(\cdot))$.
    Compute $\underline{V}_h(\cdot) = \mathbb{E}_{a\sim\underline{\mu}_h, b\sim\underline{\nu}_h}\underline{Q}_h(\cdot, a, b)$.
    Set $\overline{Q}_h(\cdot,\cdot,\cdot)$ as (6).
    Compute the NE of $\overline{Q}_h(\cdot,\cdot,\cdot)$ as $(\overline{\mu}_h(\cdot), \overline{\nu}_h(\cdot))$.
    Compute $\overline{V}_h(\cdot) = \mathbb{E}_{a\sim\overline{\mu}_h, b\sim\overline{\nu}_h}\overline{Q}_h(\cdot, a, b)$.
end for
Output $\underline{\mu} = (\underline{\mu}_1, \underline{\mu}_2, \cdots, \underline{\mu}_H)$, $\overline{\nu} = (\overline{\nu}_1, \overline{\nu}_2, \cdots, \overline{\nu}_H)$, $\{\underline{V}_h\}_{h=1}^H$, $\{\overline{V}_h\}_{h=1}^H$.
```

Algorithm 2 Pessimistic and Optimistic Value Iteration with Reference Advantage Decomposition

```
Input: Dataset $\mathcal{D} = \big\{(s_h^k, a_h^k, b_h^k, r_h^k, s_{h+1}^k)\big\}_{k,h=1}^{n,H}$. Failure Probability $\delta$.
Initialization: Randomly split the dataset $\mathcal{D}$ into $\mathcal{D}_{\text{ref}}$, $\mathcal{D}_0$, $\{\mathcal{D}_{h,1}\}_{h=1}^H$ with $|\mathcal{D}_{\text{ref}}| = n/3$, $|\mathcal{D}_0| = n/3$ $|\mathcal{D}_h| = n/(3H)$.
Set $\underline{V}_{H+1} = \overline{V}_{H+1} = 0$.
Learn the reference value function $\underline{V}_{\text{ref}}, \overline{V}_{\text{ref}} \leftarrow \text{POVI}(\mathcal{D}_{\text{ref}})$.
Set $\widehat{P}_{h,0}$ and $\widehat{r}_{h,0}$ as (2) and (3) using the dataset $\mathcal{D}_0$ for all $h \in [H]$.
Set $\widehat{P}_{h,1}$ and $\widehat{r}_{h,1}$ as (2) and (3) using the dataset $\mathcal{D}_{h,1}$ for all $h \in [H]$.
Set $\underline{b}_{h,0}$ and $\overline{b}_{h,0}$ as (9) using the dataset $\mathcal{D}_0$ for all $h \in [H]$.
for time $h = H, H-1, \dots, 1$ do
    Set $\underline{b}_{h,1}$ and $\overline{b}_{h,1}$ as (10) using the dataset $\mathcal{D}_{h,1}$ for all $h \in [H]$.
    Set $\underline{Q}_h(\cdot,\cdot,\cdot)$ as (7).
    Compute the NE of $\underline{Q}_h(\cdot,\cdot,\cdot)$ as $(\underline{\mu}_h(\cdot), \underline{\nu}_h(\cdot))$.
    Compute $\underline{V}_h(\cdot) = \mathbb{E}_{a\sim\underline{\mu}_h, b\sim\underline{\nu}_h}\underline{Q}_h(\cdot, a, b)$.
    Set $\overline{Q}_h(\cdot,\cdot,\cdot)$ as (8).
    Compute the NE of $\overline{Q}_h(\cdot,\cdot,\cdot)$ as $(\overline{\mu}_h(\cdot), \overline{\nu}_h(\cdot))$.
    Compute $\overline{V}_h(\cdot) = \mathbb{E}_{a\sim\overline{\mu}_h, b\sim\overline{\nu}_h}\overline{Q}_h(\cdot, a, b)$.
end for
Output: $\underline{\mu} = (\underline{\mu}_1, \underline{\mu}_2, \cdots, \underline{\mu}_H)$, $\overline{\nu} = (\overline{\nu}_1, \overline{\nu}_2, \cdots, \overline{\nu}_H)$.
```

B Related Work

Here we focus on the theoretical work on two-player zero-sum Markov games and offline RL.

Two-player zero-sum Markov game. Zero-sum Markov game has been widely studied since the seminal work (Shapley, 1953). When the transition kernel is unknown, different sampling oracles are utilized to acquire samples, including online sampling (Bai & Jin, 2020; Xie et al., 2020a; Liu et al., 2021; Bai et al., 2020; Jin et al., 2021a; Song et al., 2021), generative model sampling (Sidford et al., 2020; Cui & Yang, 2020; Zhang et al., 2020; Jia et al., 2019). For offline sampling oracle, Zhang et al. (2021b) provides finite sample bound for a decentralized algorithm with network communication under uniform concentration assumption and Abe & Kaneko (2020) considers offline policy evaluation, again under the uniform concentration assumption. None of these works considers the minimal dataset assumption that allows NE learning in zero-sum Markov games.

Offline single-agent RL. Theoretical analysis of offline RL can be traced back to Szepesvári & Munos (2005), under the uniform concentration assumption (analogue to Assumption 2.3). This assumption has been extensively investigated (Xie & Jiang, 2021; Xie et al., 2020b; Yin et al., 2020; 2021; Ren et al., 2021). Recently, a line of works showed that the pessimism principle allows offline policy optimization under a much weaker assumption, single policy concentration, both in tabular case and with function approximation (Rashidinejad et al., 2021; Yin & Wang, 2021; Xie et al., 2021; Jin et al., 2021b; Uehara & Sun, 2021; Uehara et al., 2021; Zanette et al., 2021). One closely related work is Xie et al. (2021), which utilizes the reference advantage function technique and bernstein-type bonus to show a minimax sample complexity $\widetilde{O}(SC^*H^3/n)$ in finite-horizon MDP. We show that the counterpart of single policy concentration in zero-sum Markov game is insufficient for NE strategy learning and use the pessimism principle to design algorithm that works under the unilateral concentration assumption.

C Proof in Section 3

Proof of Theorem 3.1. We consider a bandits game with two actions for each player here. The action set is $\mathcal{A} = \{a_1, a_2\}$ for the first (max) player and $\mathcal{B} = \{b_1, b_2\}$ for the second (min) player. We construct the following two bandit games with deterministic rewards. Then the unique NE of the first bandits game is (a_1, b_1) and the unique NE of the second bandits game is (a_2, b_2). Now we set the exploration strategy ρ to be

$$\begin{array}{ll} r(a_1,b_1)=0.5 & r(a_1,b_2)=1 \\ r(a_2,b_1)=0 & r(a_2,b_2)=0.5 \end{array}$$

Bandits Game 1

$$\begin{array}{ll} r(a_1,b_1)=0.5 & r(a_1,b_2)=0 \\ r(a_2,b_1)=1 & r(a_2,b_2)=0.5 \end{array}$$

Bandits Game 2

$$\rho(a,b) = \begin{cases} 1/2 & (a,b) = (a_1,b_1) \text{ or } (a,b) = (a_2,b_2), \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to verify that both bandits games with exploration strategy ρ is in the class $\mathcal{X}$. However, as the dataset only provides information about (a_1, b_1) and (a_2, b_2), which have the same reward, it is impossible for any algorithm to distinguish between these two bandits games. Suppose the output of $\mathbf{ALG}$ is $\pi = (\mu, \nu)$ with $\mu(a_1) = p$, $\mu(a_2) = 1 - p$ and $\nu(b_1) = q$, $\nu(b_2) = 1 - q$, then we have that π is an $0.5(2 - p - q)$-NE for the first bandits game and $0.5(p + q)$-NE for the second bandits game. As a result, there exists a bandits game such that $\mathbf{ALG}$ only outputs an at most 0.5-approximate NE strategy. □

Proof of Theorem 3.4. Similar to the proof of Theorem 3.1, we construct the following two bandits games with deterministic rewards. Then the (unique) NE of the first bandits game is (a_1, b_1) and the

$$\begin{array}{ll} r(a_1,b_1)=0.25 & r(a_1,b_2)=0.5 \\ r(a_2,b_1)=0 & r(a_2,b_2)=0.75 \end{array}$$

Bandits Game 3

$$\begin{array}{ll} r(a_1,b_1)=0.25 & r(a_1,b_2)=0.5 \\ r(a_2,b_1)=1 & r(a_2,b_2)=0.75 \end{array}$$

Bandits Game 4

(unique) NE of the second bandits game is (a_2, b_2). Now we set the exploration strategy ρ to be

$$\rho(a,b) = \begin{cases} 0 & (a,b) = (a_2, b_1), \\ 1/3 & \text{otherwise.} \end{cases}$$

It is straightforward to verify that both bandits game with exploration strategy ρ is in the class defined in Theorem 3.1. The NE for the first bandits game is (a_1, b_1) and the NE for the second bandits game is (a_2, b_2). The dataset contains data on (a_1, b_1), (a_1, b_2), (a_2, b_2) and no data on (a_2, b_1). It is impossible for an algorithm to distinguish between these two bandits games as they are consistent on the given dataset and they all satisfy the dataset coverage assumption that only one action is not covered. The rest of the proof follows the proof of Theorem 3.1 and with some computation we can show that the output of **ALG** must be a 0.25-approximate NE for one of the bandits games. □

D Proof in Section 4.1

Lemma D.1. *(Concentration) With probability* $1-\delta$*, we have*

$$\left| r_h(s,a,b) - \widehat{r}_h(s,a,b) + \left\langle P_h(\cdot|s,a,b) - \widehat{P}_h(\cdot|s,a,b), \underline{V}_{h+1}(\cdot) \right\rangle \right| \le \underline{b}_h(s,a,b),$$

$$\left| r_h(s,a,b) - \widehat{r}_h(s,a,b) + \left\langle P_h(\cdot|s,a,b) - \widehat{P}_h(\cdot|s,a,b), \overline{V}_{h+1}(\cdot) \right\rangle \right| \le \underline{b}_h(s,a,b),$$

$$\frac{1}{n_h(s,a,b) \vee 1} \le \frac{8H\iota}{nd_h^\rho(s,a,b)}.$$

holds for all $h \in [H]$*,* $s \in \mathcal{S}$*,* $a \in \mathcal{A}$ *and* $b \in \mathcal{B}$*. We define this as the good event* $\mathcal{G}$*.*

Proof. We provide the proof for the first argument and the proof for the second argument holds similarly. For all s, a, b, h, we have

$$|r_h(s,a,b) - \widehat{r}_h(s,a,b)| \le H\sqrt{\frac{1}{n_h(s,a,b) \vee 1}},$$

as whenever $n_h(s,a,b) \ge 1$, $\widehat{r}_h(s,a,b) = r_h(s,a,b)$. For the concentration on $\left\langle \widehat{P}(\cdot|s,a,b), \underline{V}_{h+1}(\cdot) \right\rangle$, note that $\underline{V}_{h+1}$ only depends on the dataset $\{\mathcal{D}_t\}_{t=h+1}^H$ while $\widehat{P}_h(\cdot|s,a,b)$ only depends on the dataset $\mathcal{D}_h$, which means they are independent and then Hoeffding's inequality can be applied:

$$\left\langle P_h(\cdot|s,a,b) - \widehat{P}_h(\cdot|s,a,b), \underline{V}_{h+1}(\cdot) \right\rangle \le 2\sqrt{\frac{H^2\iota}{n_h(s,a,b) \vee 1}}.$$

The second argument holds similarly. For the third argument, the proof is from Lemma B.1 in Xie et al. (2021). □

Lemma D.2. *(Pessimism) Under the good event* $\mathcal{G}$*, we have that* $\underline{V}_h(s) \le V_h^{\mu,*}(s)$ *and* $\overline{V}_h(s) \ge V_h^{*,\overline{\nu}}(s)$ *hold for all* $h \in [H]$ *and* $s \in \mathcal{S}$*.*

Proof. We prove this lemma by induction. The inequalities trivially hold for $h = H+1$. If the inequalities hold for timestep $h+1$, now we consider timestep h. By the definition of $\overline{Q}_h(s,a,b)$, we have

$$\begin{aligned}
\underline{Q}_h(s,a,b) &= \left(\widehat{r}_h(s,a,b) + (\widehat{P}_h \cdot \underline{V}_{h+1})(s,a,b) - \underline{b}_h(s,a,b)\right) \vee 0 \\
&\le \left(r(s,a,b) + (P \cdot V_{h+1}^{\mu,*})(s,a,b)\right) \vee 0 \\
&= r(s,a,b) + (P \cdot V_{h+1}^{\mu,*})(s,a,b) \\
&= Q_h^{\mu,*}(s,a,b),
\end{aligned}$$

where the inequality is from Lemma D.1. With the pessimism on the state-action value function, we can prove the pessimism on the state value function.

$$\begin{aligned}\underline{V}_h(s) =&\mathbb{E}_{\underline{\mu}_h,\underline{\nu}_h}\underline{Q}_h(s,a,b)\\ \leq&\mathbb{E}_{\underline{\mu}_h,\mathrm{br}(\underline{\mu}_h)}\underline{Q}_h(s,a,b)\\ \leq&\mathbb{E}_{\underline{\mu}_h,\mathrm{br}(\underline{\mu}_h)}Q_h^{\underline{\mu},*}(s,a,b)\\ =&V_h^{\underline{\mu},*}(s,a,b),\end{aligned}$$

where the first inequality is from the definition of NE and the second inequality is from the pessimism of the state-action value function. The arguments for $\overline{V}_h$ hold similarly. Then by mathematical induction we can prove the lemma. □

Lemma D.3. *Under the good event $\mathcal{G}$, for all $h \in [H]$ and $s_h \in \mathcal{S}$, we have*

$$V_h^*(s_h) - V_h^{\underline{\mu},*}(s_h) \leq V_h^{\mu^*,\underline{\nu}}(s_h) - \underline{V}_h(s_h) \leq 2\mathbb{E}_{\mu^*,\underline{\nu}}\left[\sum_{t=h}^{H}\underline{b}_t(s_t,a_t,b_t)|s_h\right],$$

$$V_h^{*,\overline{\nu}}(s_h) - V_h^*(s_h) \leq \overline{V}_h(s_h) - V_h^{\overline{\mu},\nu^*}(s_h) \leq 2\mathbb{E}_{\overline{\mu},\nu^*}\left[\sum_{t=h}^{H}\overline{b}_t(s_t,a_t,b_t)|s_h\right].$$

Proof. We prove the first argument and the second argument can be proven similarly. By the definition of NE, we have $V_h^* \leq V_h^{\mu^*,\underline{\nu}}$. Combined with Lemma D.2, we have the first inequality. For the second inequality, we have

$$\begin{aligned}&V_h^{\mu^*,\underline{\nu}}(s_h) - \underline{V}_h(s_h)\\ =&\mathbb{E}_{\mu_h^*,\underline{\nu}_h}Q_h^{\mu^*,\underline{\nu}}(s_h,a_h,b_h) - \mathbb{E}_{\underline{\mu}_h,\underline{\nu}_h}\underline{Q}_h(s_h,a_h,b_h)\\ \leq&\mathbb{E}_{\mu_h^*,\underline{\nu}_h}Q_h^{\mu^*,\underline{\nu}}(s_h,a_h,b_h) - \mathbb{E}_{\mu_h^*,\underline{\nu}_h}\underline{Q}_h(s_h,a_h,b_h)\\ =&\mathbb{E}_{\mu_h^*,\underline{\nu}_h}\left[Q_h^{\mu^*,\underline{\nu}}(s_h,a_h,b_h) - \underline{Q}(s_h,a_h,b_h)\right]\\ =&\mathbb{E}_{\mu_h^*,\underline{\nu}_h}\left[r_h(s_h,a_h,b_h) + \left\langle P_h(\cdot|s_h,a_h,b_h), V_{h+1}^{\mu^*,\underline{\nu}}(\cdot)\right\rangle - \widehat{r}_h(s_h,a_h,b_h) - \left\langle \widehat{P}_h(\cdot|s_h,a_h,b_h), \underline{V}_{h+1}(\cdot)\right\rangle + \underline{b}_h(s_h,a_h,b_h)\right]\\ \leq&\mathbb{E}_{\mu_h^*,\underline{\nu}_h}\left[\left\langle P_h(\cdot|s_h,a_h,b_h), V_{h+1}^{\mu^*,\underline{\nu}}(\cdot) - \underline{V}_{h+1}(\cdot)\right\rangle + 2\underline{b}_h(s_h,a_h,b_h)\right] && \text{(Lemma D.1)}\\ =&\mathbb{E}_{\mu_h^*,\underline{\nu}_h}\left[V_{h+1}^*(s_{h+1}) - \underline{V}_{h+1}^*(s_{h+1})|s_h\right] + 2\mathbb{E}_{\mu_h^*,\underline{\nu}_h^*}\underline{b}_h(s_h,a_h,b_h)\\ \leq&2\mathbb{E}_{\mu^*,\underline{\nu}}\left[\sum_{t=h}^{H}\underline{b}_h(s_t,a_t,b_t)|s_h\right].\end{aligned}$$

□

Theorem D.4. *Suppose Assumption 2.2 holds. For any $0 < \delta < 1$, with probability $1-\delta$, the output policy $\pi = (\underline{\mu}, \overline{\nu})$ of Algorithm 1 satisfies*

$$V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) \leq 64\sqrt{\frac{C^*SABH^5\iota^2}{n}},$$

$$V_1^{*,\overline{\nu}}(s_1) - V_1^*(s_1) \leq 64\sqrt{\frac{C^*SABH^5\iota^2}{n}}.$$

As a result, we have

$$\mathrm{Gap}(\underline{\mu},\overline{\nu}) \leq \widetilde{O}\left(\frac{C^*SABH^5}{n}\right).$$

Proof. By Lemma D.3, with probability at least $1-\delta$, we have

$$\begin{aligned}
& V_1^{\mu^*,*}(s_1) - V_1^{\underline{\mu},*}(s_1) \\
\le & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \underline{b}_h(\mathrm{s_h}, \mathrm{a_h}, \mathrm{b_h}) \\
= & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[4\sqrt{\frac{H^2\iota}{n_h(s,a,b)\vee 1}} \right] \\
\le & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[32\sqrt{\frac{H^3\iota^2}{n d_h^{\rho}(s,a,b)}} \right] && \text{(Lemma D.1)} \\
= & 2\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) \left[32\sqrt{\frac{H^3\iota^2}{n d_h^{\rho}(s,a,b)}} \right] \\
\le & 64\sum_{h=1}^{H} \sum_{(s,a,b)} \left[\sqrt{\frac{d_h^{\mu^*,\underline{\nu}}(s,a,b) C^* H^3 \iota^2}{n}} \right] \\
\le & 64\sqrt{SABH} \cdot \sqrt{\frac{\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) C^* H^3 \iota^2}{n}} && \text{(Cauchy-Schwarz Inequality)} \\
= & 64\sqrt{\frac{C^* SABH^5 \iota^2}{n}}.
\end{aligned}$$

Similarly we have

$$V_1^{*,\overline{\nu}}(s_1) - V_1^{*}(s_1) \le 64\sqrt{\frac{C^* SABH^5 \iota^2}{n}}.$$

As a result, we have

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \le V_1^{*,\overline{\nu}}(s_1) - V_1^{*}(s_1) + V_1^{\mu^*,*}(s_1) - V_1^{\underline{\mu},*}(s_1) \le \widetilde{O}\left(\frac{C^* SABH^5}{n} \right).$$

□

Theorem D.5. *Suppose Assumption 2.2 holds. For any $0 < \delta < 1$ and strategy μ, ν, with probability $1-\delta$, the pessimistic value $\underline{V}_h$ and optimistic estimate $\overline{V}_h$ of Algorithm 1 satisfies*

$$\mathbb{E}_{\mu^*,\nu}\left[V_h^*(s_h) - \underline{V}_h(s_h) \right] \le 64\sqrt{\frac{C^* SABH^5 \iota^2}{n}},$$

$$\mathbb{E}_{\mu,\nu^*}\left[\overline{V}_h(s_h) - V_h^*(s_h) \right] \le 64\sqrt{\frac{C^* SABH^5 \iota^2}{n}},$$

where s_h is sampled from the trajectory following the strategy in the expectation at timestep h.

Proof. We prove the first argument and the second argument can be proven similarly. By Lemma D.3, under good event $\mathcal{G}$ for all state s we have

$$\begin{aligned}
& V_h^*(s) - V_h^{\underline{\mu},*}(s) \\
\le & 2\sum_{t=h}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[\underline{b}_h(s_t, a_t, b_t) | s_h = s \right].
\end{aligned}$$

We define $\nu' = (\nu_1, \cdots, \nu_{h-1}, \underline{\nu}_h, \cdots, \underline{\nu}_H)$. Then we have

$$\mathbb{E}_{\mu^*,\nu}\left[V_h^*(s_h) - \underline{V}_h(s_h) \right] \le \mathbb{E}_{\mu^*,\nu} \left[2\sum_{t=h}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[\underline{b}_h(s_t, a_t, b_t) | s_h = s \right] | s \right]$$

$$=2\sum_{t=h}^{H}\mathbb{E}_{\mu^*,\nu'}\left[\underline{b}_h(s_t,a_t,b_t)\right].$$

Then following the proof of Theorem D.4, we can prove the argument.

□

E Proof in Section 4.2

For simplicity, we only provide the guarantee for the max player and the guarantee for the min player can be proven in a similar manner.

Lemma E.1. *(Concentration) There exists some absolute constant $c > 0$ such that the concentration event $\mathcal{G}'$ holds with probability at least $1-\delta$, where*

$$\left|\widehat{r}_{h,0}(s,a,b) - r_{h,0}(s,a,b) + \left[\left(\widehat{P}_{h,0} - P_h\right)\underline{V}^{\text{ref}}_{h+1}\right](s,a,b)\right| \le c\left(\sqrt{\frac{\operatorname{Var}_{\widehat{P}_{h,0}(s,a,b)}(\underline{V}^{\text{ref}}_{h+1})\iota}{n_{h,0}(s,a,b)\vee 1}} + \frac{H\iota}{n_{h,0}(s,a,b)\vee 1}\right),$$

$$\left|\left[\left(\widehat{P}_{h,1} - P_h\right)\left(\underline{V}_{h+1} - \underline{V}^{\text{ref}}_{h+1}\right)\right](s,a,b)\right| \le c\left(\sqrt{\frac{\operatorname{Var}_{\widehat{P}_{h,1}(s,a,b)}(\underline{V}_{h+1} - \underline{V}^{\text{ref}}_{h+1})\iota}{n_{h,1}(s,a,b)\vee 1}} + \frac{H\iota}{n_{h,1}(s,a,b)\vee 1}\right),$$

$$\frac{1}{n_{h,0}(s,a,b)\vee 1} \le c\frac{\iota}{nd^{\rho}_{h}(s,a,b)}, \frac{1}{n_{h,1}(s,a,b)\vee 1} \le c\frac{H\iota}{nd^{\rho}_{h}(s,a,b)}$$

Proof. The proof is a direct application of Lemma C.1 in Xie et al. (2021) with s, a replaced by s, a, b. □

Lemma E.2. *For all $h \in [H]$ and $s \in \mathcal{S}$, we have $\underline{V}_h(s) \ge \underline{V}^{\text{ref}}_h(s)$.*

Proof. By the update rule (7), we have $\underline{Q}_h(s,a,b) \ge \underline{Q}^{\text{ref}}_h(s,a,b)$ for $h \in [H]$ and $s,a,b \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$. Then by the definition of NE, we have

$$\underline{V}_h(s) = \mathbb{E}_{\underline{\mu}_h,\underline{\nu}_h}\underline{Q}_h(s,a,b) \ge \mathbb{E}_{\underline{\mu}^{\text{ref}}_h,\underline{\nu}_h}\underline{Q}_h(s,a,b) \ge \mathbb{E}_{\underline{\mu}^{\text{ref}}_h,\underline{\nu}_h}\underline{Q}^{\text{ref}}_h(s,a,b) \ge \mathbb{E}_{\underline{\mu}^{\text{ref}}_h,\underline{\nu}^{\text{ref}}_h}\underline{Q}^{\text{ref}}_h(s,a,b) = \underline{V}^{\text{ref}}_h(s).$$

□

Lemma E.3. *(Pessimism) Under the good event $\mathcal{G}'$, we have that $\underline{V}_h(s) \le V^{\mu,*}_h(s)$ holds for all $h \in [H]$ and $s \in \mathcal{S}$.*

Proof. We prove this lemma by induction. The inequalities trivially hold for $h = H+1$. If the inequalities hold for $h+1$, now we consider h.

$$\begin{aligned}
&\underline{Q}_h(s,a,b)\\
=&\left\{\widehat{r}_{h,0}(s,a,b) + (\widehat{P}_{h,0}\cdot\underline{V}^{\text{ref}}_2)(s,a,b) - \underline{b}_{h,0}(s,a,b) + (\widehat{P}_{h,1}\cdot(\underline{V}_{h+1} - \underline{V}^{\text{ref}}_{h+1}))(s,a,b) - \underline{b}_{h,1}(s,a,b)\right\} \vee \underline{Q}^{\text{ref}}_h(s,a,b)\\
\le&\max\left\{r_h(s,a,b) + (P_h\cdot\underline{V}^{\text{ref}}_{h+1})(s,a,b) + \left(P_h\cdot\left(\underline{V}_{h+1} - \underline{V}^{\text{ref}}_{h+1}\right)\right)(s,a,b), \underline{Q}^{\text{ref}}_h(s,a,b)\right\}\\
=&\max\left\{r_h(s,a,b) + (P_h\cdot\underline{V}_{h+1})(s,a,b), \underline{Q}^{\text{ref}}_h(s,a,b)\right\}\\
\le&\max\left\{r_h(s,a,b) + (P_h\cdot\underline{V}_{h+1})(s,a,b), r_h(s,a,b) + (P_h\cdot\underline{V}^{\text{ref}}_{h+1})(s,a,b)\right\} && \text{(Lemma D.2)}\\
\le& r_h(s,a,b) + (P_h\cdot\underline{V}_{h+1})(s,a,b) && \text{(Lemma E.2)}\\
\le& r_h(s,a,b) + (P_h\cdot V^{\mu,*}_{h+1})(s,a,b) && \text{(Induction hypothesis)}\\
=& Q^{\mu,*}_h(s,a,b).
\end{aligned}$$

Then by the definition of NE, we have

$$\begin{aligned}\underline{V}_h(s) =&\mathbb{E}_{\underline{\mu}_h,\underline{\nu}_h}\underline{Q}_h(s,a,b)\\ \le&\mathbb{E}_{\underline{\mu}_h,\mathrm{br}(\underline{\mu}_h)}\underline{Q}_h(s,a,b)\\ \le&\mathbb{E}_{\underline{\mu}_h,\mathrm{br}(\underline{\mu}_h)}Q_h^{\underline{\mu},*}(s,a,b)\\ =&V_h^{\dagger\underline{\mu},*}(s).\end{aligned}$$

With mathematical induction we can prove the lemma. □

Lemma E.4. *Under the good event $\mathcal{G}'$, we have*

$$V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1)\le 2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h)+2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,1}(s_h,a_h,b_h)$$

Proof.

$$\begin{aligned}&V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1)\\ =&\mathbb{E}_{\mu_1^*,\underline{\nu}_1}Q_1^{\mu^*,\underline{\nu}}(s_1,a_1,b_1)-\mathbb{E}_{\underline{\mu}_1,\underline{\nu}_1}\underline{Q}_1(s_1,a_1,b_1)\\ \le&\mathbb{E}_{\mu_1^*,\underline{\nu}_1}Q_1^{\mu^*,\underline{\nu}}(s_1,a_1,b_1)-\mathbb{E}_{\mu_1^*,\underline{\nu}_1}\underline{Q}_1(s_1,a_1,b_1)\\ =&\mathbb{E}_{\mu_1^*,\underline{\nu}_1}\left[Q_1^{\mu^*,\underline{\nu}}(s_1,a_1,b_1)-\underline{Q}_1(s_1,a_1,b_1)\right]\\ =&\mathbb{E}_{\mu_1^*,\underline{\nu}_1}[r_1(s_1,a_1,b_1)+\left\langle P_1(\cdot|s_1,a_1,b_1),V_2^{\mu^*,\underline{\nu}}(\cdot)\right\rangle-\underline{V}_1^{\mathrm{ref}}(s_1)\vee\\ &\left\{\widehat{r}_{1,0}(s_1,a_1,b_1)+(\widehat{P}_{1,0}\underline{V}_2^{\mathrm{ref}})(s_1,a_1,b_1)-\underline{b}_{1,0}(s_1,a_1,b_1)+(\widehat{P}_{1,1}(\underline{V}_2-\underline{V}_2^{\mathrm{ref}}))(s_1,a_1,b_1)-\underline{b}_{1,1}(s_1,a_1,b_1)\right\}]\\ \le&\mathbb{E}_{\mu_1^*,\underline{\nu}_1}\left[\left\langle P_1(\cdot|s_1,a_1,b_1),V_2^{\mu^*,\underline{\nu}}(\cdot)-\underline{V}_2(\cdot)\right\rangle+2\underline{b}_{1,0}(s_1,a_1,b_1)+2\underline{b}_{1,1}(s_1,a_1,b_1)\right]\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Lemma E.1)}\\ =&\mathbb{E}_{\mu_1^*,\underline{\nu}_1}\left[V_2^{\mu^*,\underline{\nu}}(s_2)-\underline{V}_2^*(s_2)\right]+2\mathbb{E}_{\mu_1^*,\underline{\nu}_1^*}\underline{b}_{1,0}(s_1,a_1,b_1)+2\mathbb{E}_{\mu_1^*,\underline{\nu}_1^*}\underline{b}_{1,1}(s_1,a_1,b_1)\\ \le&2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h)+2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,1}(s_h,a_h,b_h),\end{aligned}$$

where the last inequality is from telescoping the timestep H. □

Lemma E.5. *For any strategy ν, we have*

$$\sum_{h=1}^{H}\sum_{(s,a,b)}d_h^{\mu^*,\nu}(s,a,b)\operatorname*{Var}_{P_h(s,a,b)}(V_{h+1}^{\mu^*,\nu})\le H^2$$

Proof.

$$\begin{aligned}\sum_{h=1}^{H}\sum_{(s,a,b)}d_h^{\mu^*,\nu}(s,a,b)\operatorname*{Var}_{P_h(s,a,b)}(V_h^{\mu^*,\nu})&=\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\nu}\left[\mathrm{Var}\left[V_{h+1}^*(s_{h+1})|s_h,a_h,b_h\right]\right]\\ &=\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\nu}\left[\mathbb{E}\left[\left(V_{h+1}^*(s_{h+1})+r_h(s_h,a_h,b_h)-V_h^*(s_h)\right)^2|s_h,a_h,b_h\right]\right]\\ &=\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\nu}\left[\left(V_{h+1}^*(s_{h+1})+r_h(s_h,a_h,b_h)-V_h^*(s_h)\right)^2\right]\\ &=\mathbb{E}_{\mu^*,\nu}\left[\left(\sum_{h=1}^{H}\left(V_{h+1}^*(s_{h+1})+r_h(s_h,a_h,b_h)-V_h^*(s_h)\right)\right)^2\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\mu^*,\nu}\left[\left(\sum_{h=1}^{H} r_h(s_h,a_h,b_h) - V_1^*(s_1)\right)^2\right] \\
&= \operatorname*{Var}_{\mu^*,\nu}\left(\sum_{h=1}^{H} r_h(s_h,a_h,b_h)\right) \\
&\le H^2.
\end{aligned}
$$

□

Lemma E.6. *The output strategy $\pi = (\underline{\mu}, \overline{\nu})$ and the pessimistic estimate $\underline{V}$ of Algorithm 1 satisfy*

$$
V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1) \ge \mathbb{E}_{\mu^*,\underline{\nu}}\left[V_h^{\mu^*,\underline{\nu}}(s_h) - \underline{V}_h(s_h)\right].
$$

Proof. We prove the argument for $h = 2$ first.

$$
\begin{aligned}
&V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1) \\
\ge& \mathbb{E}_{\mu^*,\underline{\nu}}[Q_1^{\mu^*,\underline{\nu}}(s_1,a_1,b_1) - \underline{Q}_1(s_1,a_1,b_1)] \\
\ge& \mathbb{E}_{\mu^*,\underline{\nu}}\left[r_1(s_1,a_1,b_1) + \left\langle P_1(\cdot|s_1,a_1,b_1), V_2^{\mu^*,\underline{\nu}}(\cdot)\right\rangle\right] \\
-& \mathbb{E}_{\mu^*,\underline{\nu}}\left[\widehat{r}_{1,0}(s_1,a_1,b_1) + (\widehat{P}_{1,0}\underline{V}_2^{\mathrm{ref}})(s_1,a_1,b_1) - \underline{b}_{1,0}(s_1,a_1,b_1) + (\widehat{P}_{1,1}(\underline{V}_2 - \underline{V}_2^{\mathrm{ref}}))(s_1,a_1,b_1) - \underline{b}_{1,1}(s_1,a_1,b_1)\right] \\
\ge& \mathbb{E}_{\mu^*,\underline{\nu}}\left[r_1(s_1,a_1,b_1) + \left\langle P_1(\cdot|s_1,a_1,b_1), V_2^{\mu^*,\underline{\nu}}(\cdot)\right\rangle\right] - \mathbb{E}_{\mu^*,\underline{\nu}}\left[r_1(s_1,a_1,b_1) + \langle P_1(\cdot|s_1,a_1,b_1), \underline{V}_2(\cdot)\rangle\right] \\
=& \mathbb{E}_{\mu^*,\underline{\nu}}\left[V_2^{\mu^*,\underline{\nu}}(s_2) - \underline{V}_2(s_2)\right].
\end{aligned}
$$

We can prove the lemma for arbitrary h by telescoping the argument to timestep h.

□

Lemma E.7. *For $n \ge \widetilde{O}(\sqrt{C^*SABH^3})$, we have*

$$
\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h) \le \widetilde{O}\left(\sqrt{\frac{C^*SABH^3}{n}}\sqrt{V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1)}\right) + \widetilde{O}\left(\sqrt{\frac{C^*SABH^3}{n}}\right).
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h) \\
=& c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,0}(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,0}(s,a,b)\vee 1}} + \frac{H\iota}{n_{h,0}(s,a,b)\vee 1}\right) \\
\le& c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{c\,\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{nd_h^{\rho}(s,a,b)}} + \frac{cH\iota}{nd_h^{\rho}(s,a,b)} + \frac{cH\iota}{nd_h^{\rho}(s,a,b)}\right) \\
=& c^2\sum_{h=1}^{H}\sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)\left(\sqrt{\frac{\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{nd_h^{\rho}(s,a,b)}} + \frac{H\iota}{nd_h^{\rho}(s,a,b)}\right) \\
\le& c^2\sum_{h=1}^{H}\sum_{(s,a,b)}\left(\sqrt{\frac{C^*d_h^{\mu^*,\underline{\nu}}(s,a,b)\,\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n}} + \frac{C^*H\iota}{n}\right) \\
\le& c^2\sqrt{SABH}\cdot\sqrt{\frac{C^*\iota\sum_{h=1}^{H}\sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)\,\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})}{n}} + \frac{c^2SABC^*H\iota}{n}
\end{aligned}
$$

$$\le c^2\sqrt{C^*SABH\iota}\cdot\sqrt{\frac{\sum_{h=1}^H \mathbb{E}_{\mu^*,\underline{\nu}}\left[\mathrm{Var}_{P_h(s,a,b)}(\underline{V}^{\mathrm{ref}}_{h+1})\right]}{n}}+\frac{c^2SABC^*H\iota}{n}$$

$$\le c^2\sqrt{C^*SABH\iota}\cdot\sqrt{\frac{\sum_{h=1}^H \mathbb{E}_{\mu^*,\underline{\nu}}\left[\mathrm{Var}_{P_h(s,a,b)}(V^{\mu^*,\underline{\nu}}_{h+1})+2H[P_h(V^{\mu^*,\underline{\nu}}_{h+1}-\underline{V}^{\mathrm{ref}}_{h+1})](s,a,b)\right]}{n}}+\frac{c^2SABC^*H\iota}{n}$$

(Lemma G.4)

$$\le c^2\sqrt{C^*SABH\iota}\cdot\sqrt{\frac{H^2+2H\sum_{h=1}^H \mathbb{E}_{\mu^*,\underline{\nu}}\left[V^{\mu^*,\underline{\nu}}_{h+1}(s_{h+1})-\underline{V}^{\mathrm{ref}}_{h+1}(s_{h+1})\right]}{n}}+\frac{c^2SABC^*H\iota}{n}$$

(Lemma E.5)

$$=c^2\sqrt{C^*SABH\iota}\cdot\sqrt{\frac{H^2+2H\sum_{h=1}^H \mathbb{E}_{\mu^*,\underline{\nu}}\left[V^{\mu^*,\underline{\nu}}_{h+1}(s_{h+1})-V^*_{h+1}(s_{h+1})+V^*_{h+1}(s_{h+1})-\underline{V}^{\mathrm{ref}}_{h+1}(s_{h+1})\right]}{n}}$$
$$+\frac{c^2SABC^*H\iota}{n}$$

$$\le c^2\sqrt{C^*SABH\iota}\cdot\sqrt{\frac{H^2+2H^2(V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1))+128H\sqrt{\frac{C^*SABH^5\iota^2}{n_{\mathrm{ref}}}}}{n}}+\frac{c^2SABC^*H\iota}{n}$$

(Lemma E.6 and Theorem D.5)

$$\le\frac{c^2\sqrt{C^*SABH^3\iota}}{\sqrt{n}}+\frac{c^2\sqrt{384C^*SABH^2\iota\sqrt{C^*SABH^5\iota^2}}}{n^{3/4}}+\frac{c\sqrt{2C^*SABH^3\iota}}{\sqrt{n}}\sqrt{V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1)}$$
$$+\frac{c^2SABC^*H\iota}{n}.$$

□

Lemma E.8. *For $n\ge\widetilde{O}(\sqrt{C^*SABH^4})$, we have*

$$\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^H\underline{b}_{h,1}(s_h,a_h,b_h)\le\widetilde{O}\left(\sqrt{\frac{C^*SABH^3}{n}}\right).$$

Proof.

$$\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^H\underline{b}_{h,1}(s_h,a_h,b_h)$$

$$=c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^H\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,0}(s,a,b)}(\underline{V}_{h+1}-\underline{V}^{\mathrm{ref}}_{h+1})\iota}{n_{h,1}(s,a,b)\vee 1}}+\frac{H\iota}{n_{h,1}(s,a,b)\vee 1}\right)$$

$$\le c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^H\left(\sqrt{\frac{cH\,\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}-\underline{V}^{\mathrm{ref}}_{h+1})\iota}{nd_h^\rho(s,a,b)}}+\frac{cH^2\iota}{nd_h^\rho(s,a,b)}+\frac{cH^2\iota}{nd_h^\rho(s,a,b)}\right)$$

$$\le c^2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^H\left(\sqrt{\frac{H\left[P_h(\underline{V}_{h+1}-\underline{V}^{\mathrm{ref}}_{h+1})^2\right](s,a,b)\iota}{nd_h^\rho(s,a,b)}}+\frac{H^2\iota}{nd_h^\rho(s,a,b)}\right)$$

$$=c^2\sum_{h=1}^H\sum_{(s,a,b)}d_h^{\mu^*,\underline{\nu}}(s,a,b)\left(\sqrt{\frac{H\left[P_h(\underline{V}_{h+1}-\underline{V}^{\mathrm{ref}}_{h+1})^2\right](s,a,b)\iota}{nd_h^\rho(s,a,b)}}+\frac{H^2\iota}{nd_h^\rho(s,a,b)}\right)$$

$$
\begin{aligned}
&\leq c^2 \sum_{h=1}^{H} \sum_{(s,a,b)} \left(\sqrt{\frac{C^* H d_h^{\mu^*,\underline{\nu}}(s,a,b) \left[P_h(\underline{V}_{h+1} - \underline{V}_{h+1}^{\text{ref}})^2 \right](s,a,b)\iota}{n_1}} + \frac{H^2 C^* \iota}{n_1} \right) \\
&\hfill \text{(Cauchy-Schwarz Inequality)} \\
&\leq c^2 \sqrt{SABH\iota} \sqrt{\frac{C^* H \sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) \left[P_h(\underline{V}_{h+1} - \underline{V}_{h+1}^{\text{ref}})^2 \right](s,a,b)}{n}} + \frac{c^2 SABH^3 C^* \iota}{n} \\
&\leq c^2 \sqrt{SABH\iota} \sqrt{\frac{C^* H\iota \sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) \left[P_h(V_{h+1}^* - \underline{V}_{h+1}^{\text{ref}})^2 \right](s,a,b)}{n}} + \frac{c^2 C^* SABH^3 \iota}{n} \\
&\hfill (V_{h+1}^* \geq \underline{V}_{h+1} \geq \underline{V}_{h+1}^{\text{ref}}) \\
&= c^2 \sqrt{SABH\iota} \sqrt{\frac{H^2 C^* \sum_{h=1}^{H} \sum_{s} d_{h+1}^{\mu^*,\underline{\nu}}(s)(V_{h+1}^*(s) - \underline{V}_{h+1}^{\text{ref}}(s))}{n}} + \frac{c^2 C^* SABH^3 \iota}{n} \\
&\leq c^2 \sqrt{SABH\iota} \sqrt{\frac{H^2 C^* 64 \sqrt{\frac{C^* SABH^5 \iota^2}{n_{\text{ref}}}}}{n}} + \frac{c^2 SABH^3 C^* \iota}{n} \qquad \text{(Theorem D.5)} \\
&= c^2 \sqrt{\frac{192 C^* SABH^3 \iota \sqrt{C^* SABH^5 \iota^2}}{n^{3/2}}} + \frac{c^2 C^* SABH^3 \iota}{n}.
\end{aligned}
$$

□

Theorem E.9. *Suppose Assumption 2.2 holds. For any $0 < \delta < 1$ and $n \geq \widetilde{O}(\sqrt{C^* SABH^4})$, with probability $1 - \delta$, the output policy $\pi = (\underline{\mu}, \overline{\nu})$ of Algorithm 1 satisfies*

$$
V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) \leq \widetilde{O}\left(\sqrt{\frac{C^* SABH^3}{n}} \right),
$$

$$
V_1^{*,\overline{\nu}}(s_1) - V_1^*(s_1) \leq \widetilde{O}\left(\sqrt{\frac{C^* SABH^3}{n}} \right).
$$

As a result, we have

$$
\text{Gap}(\underline{\mu}, \overline{\nu}) \leq \widetilde{O}\left(\sqrt{\frac{C^* SABH^3}{n}} \right).
$$

Proof.

$$
\begin{aligned}
& V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1) \\
\leq\, & 2\mathbb{E}_{\mu^*,\underline{\nu}} \sum_{h=1}^{H} \underline{b}_{h,0}(s_h, a_h, b_h) + 2\mathbb{E}_{\mu^*,\underline{\nu}} \sum_{h=1}^{H} \underline{b}_{h,1}(s_h, a_h, b_h) \qquad \text{(Lemma E.4)} \\
\leq\, & \widetilde{O}\left(\sqrt{\frac{C^* SABH^3}{n}} \sqrt{V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1)} \right) + \widetilde{O}\left(\sqrt{\frac{C^* SABH^3}{n}} \right) \\
& \hfill \text{(Lemma E.7 and Lemma E.8)} \\
\leq\, & \widetilde{O}\left(\sqrt{\frac{C^* SABH^3}{n}} \right) + \widetilde{O}\left(\frac{C^* SABH^3}{n} \right) \qquad \text{(Lemma G.5)} \\
=\, & \widetilde{O}\left(\sqrt{\frac{C^* SABH^3}{n}} \right).
\end{aligned}
$$

By the definition of NE, we have

$$V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) \le V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1) \le \widetilde{O}\left(\sqrt{\frac{C^*SABH^3}{n}}\right).$$

The second argument can be proven in a similar manner. Combining these two argument and we can prove that

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \le \widetilde{O}\left(\sqrt{\frac{C^*SABH^3}{n}}\right).$$

□

F Proofs in Section 4.3

F.1 Uniform Coverage

Theorem F.1. *Suppose Assumption 3.2 holds and* $d_m = \min\{d_h^\rho(s,a,b) : h \in [H], (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\}$. *For any* $0 < \delta < 1$, *with probability* $1-\delta$, *the output policy* $\pi = (\underline{\mu}, \overline{\nu})$ *of Algorithm 1 satisfies*

$$V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) \le 64\sqrt{\frac{H^5\iota^2}{nd_m}},$$

$$V_1^{*,\overline{\nu}}(s_1) - V_1^*(s_1) \le 64\sqrt{\frac{H^5\iota^2}{nd_m}}.$$

As a result, we have

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \le \widetilde{O}\left(\sqrt{\frac{H^5}{nd_m}}\right)$$

Proof. By Lemma D.3, with probability $1-\delta$ we have

$$\begin{aligned}
& V_1^{\mu^*,*}(s_1) - V_1^{\underline{\mu},*}(s_1) \\
\le & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}}\underline{b}_h(\mathrm{s_h}, \mathrm{a_h}, \mathrm{b_h}) \\
= & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}}\left[4\sqrt{\frac{H^2\iota}{n_h(s,a,b)\vee 1}}\right] \\
\le & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}}\left[32\sqrt{\frac{H^3\iota^2}{nd_h^\rho(s,a,b)}}\right] && \text{(Lemma D.1)} \\
= & 2\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)\left[32\sqrt{\frac{H^3\iota^2}{nd_h^\rho(s,a,b)}}\right] \\
\le & 64\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)\left[\sqrt{\frac{H^3\iota^2}{nd_m}}\right] \\
\le & 64\sqrt{\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)} \cdot \sqrt{\frac{\sum_{h=1}^{H}\sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)C^*H^3\iota^2}{nd_m}} \\
& && \text{(Cauchy-Schwarz Inequality)} \\
= & \sqrt{H}\cdot\sqrt{\frac{H^4\iota^2}{nd_m}}
\end{aligned}$$

$$=64\sqrt{\frac{H^5\iota^2}{nd_m}}.$$

□

Theorem F.2. *Suppose Assumption 3.2 holds and* $d_m = \min\{d_h^\rho(s,a,b) : h \in [H], (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\}$*. For any* $0 < \delta < 1$ *and strategy* μ, ν*, with probability* $1-\delta$*, the pessimistic value* $\underline{V}_h$ *and optimistic estimate* $\overline{V}_h$ *of Algorithm 1 satisfies*

$$\mathbb{E}_{\mu^*,\nu}\left[V_h^*(s_h) - \underline{V}_h(s_h)\right] \le 64\sqrt{\frac{H^5\iota^2}{nd_m}},$$

$$\mathbb{E}_{\mu,\nu^*}\left[\overline{V}_h(s_h) - V_h^*(s_h)\right] \le 64\sqrt{\frac{H^5\iota^2}{nd_m}},$$

where s_h *is sampled from the trajectory following the strategy in the expectation at timestep* h*.*

Proof. By Lemma D.3, under good event $\mathcal{G}$ for all state s we have

$$\begin{aligned}&V_h^*(s) - V_h^{\mu,*}(s)\\ \le&2\sum_{t=h}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[\underline{b}_h(s_t,a_t,b_t)|s_h = s\right]\end{aligned}$$

We define $\nu' = (\nu_1, \cdots, \nu_{h-1}, \underline{\nu}_h, \cdots, \underline{\nu}_H)$. Then we have

$$\begin{aligned}\mathbb{E}_{\mu^*,\nu}\left[V_h^*(s_h) - \underline{V}_h(s_h)\right] \le&\mathbb{E}_{\mu^*,\nu}\left[2\sum_{t=h}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[\underline{b}_h(s_t,a_t,b_t)|s_h = s\right]|s\right]\\ =&2\sum_{t=h}^{H}\mathbb{E}_{\mu^*,\nu'}\left[\underline{b}_h(s_t,a_t,b_t)\right].\end{aligned}$$

Then following the proof of Theorem F.1, we can prove the argument.

□

Lemma F.3. *Suppose Assumption 3.2 holds and* $d_m = \min\{d_h^\rho(s,a,b) : h \in [H], (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\}$*. For* $n \ge \widetilde{O}(\sqrt{H^3/d_m})$*, we have*

$$\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h) \le \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\sqrt{V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1)}\right) + \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right).$$

Proof.

$$\begin{aligned}&\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h)\\ =&c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,0}(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,0}(s,a,b)\vee 1}} + \frac{H\iota}{n_{h,0}(s,a,b)\vee 1}\right)\\ \le&c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{c\,\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{nd_h^\rho(s,a,b)}} + \frac{cH\iota}{nd_h^\rho(s,a,b)} + \frac{cH\iota}{nd_h^\rho(s,a,b)}\right)\\ \le&c^2\sum_{h=1}^{H}\sum_{(s,a,b)}d_h^{\mu^*,\underline{\nu}}(s,a,b)\left(\sqrt{\frac{\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{nd_m}} + \frac{H\iota}{nd_m}\right)\end{aligned}$$

$$\leq c^2 \sqrt{\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)} \left(\sqrt{\frac{\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) \operatorname{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{nd_m}} + \frac{H\iota}{nd_m} \right)$$

(Cauchy-Schwarz inequality)

$$\leq c^2 \sqrt{H} \cdot \sqrt{\frac{\iota \sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) \operatorname{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})}{nd_m}} + \frac{c^2 H\iota}{nd_m}$$

$$\leq c^2 \sqrt{H\iota} \cdot \sqrt{\frac{\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[\operatorname{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}}) \right]}{nd_m}} + \frac{c^2 H\iota}{nd_m}$$

$$\leq c^2 \sqrt{H\iota} \cdot \sqrt{\frac{\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[\operatorname{Var}_{P_h(s,a,b)}(V_{h+1}^{\mu^*,\underline{\nu}}) + 2H[P_h(V_{h+1}^{\mu^*,\underline{\nu}} - \underline{V}_{h+1}^{\mathrm{ref}})](s,a,b) \right]}{nd_m}} + \frac{c^2 H\iota}{nd_m}$$

(Lemma G.4)

$$\leq c^2 \sqrt{H\iota} \cdot \sqrt{\frac{H^2 + 2H \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[V_{h+1}^{\mu^*,\underline{\nu}}(s_{h+1}) - \underline{V}_{h+1}^{\mathrm{ref}}(s_{h+1}) \right]}{nd_m}} + \frac{c^2 H\iota}{nd_m}$$

(Lemma E.5)

$$= c^2 \sqrt{H\iota} \cdot \sqrt{\frac{H^2 + 2H \sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[V_{h+1}^{\mu^*,\underline{\nu}}(s_{h+1}) - V_{h+1}^{*}(s_{h+1}) + V_{h+1}^{*}(s_{h+1}) - \underline{V}_{h+1}^{\mathrm{ref}}(s_{h+1}) \right]}{nd_m}} + \frac{c^2 H\iota}{nd_m}$$

$$\leq c^2 \sqrt{H\iota} \cdot \sqrt{\frac{H^2 + 2H^2(V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1)) + 128H\sqrt{\frac{H^5\iota^2}{n_{\mathrm{ref}} d_m}}}{nd_m}} + \frac{c^2 H\iota}{nd_m}$$

(Lemma E.6 and Theorem F.2)

$$\leq \frac{c^2\sqrt{H^3\iota}}{\sqrt{nd_m}} + \frac{c^2\sqrt{384H^2\iota\sqrt{H^5\iota^2}}}{(nd_m)^{3/4}} + \frac{c\sqrt{2H^3\iota}}{\sqrt{nd_m}} \sqrt{V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1)} + \frac{c^2 H\iota}{nd_m}.$$

□

Lemma F.4. *For $n \geq \widetilde{O}(\sqrt{H^4/d_m})$, we have*

$$\mathbb{E}_{\mu^*,\underline{\nu}} \sum_{h=1}^{H} \underline{b}_{h,1}(s_h,a_h,b_h) \leq \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}} \right).$$

Proof.

$$\mathbb{E}_{\mu^*,\underline{\nu}} \sum_{h=1}^{H} \underline{b}_{h,1}(s_h,a_h,b_h)$$

$$= c\mathbb{E}_{\mu^*,\underline{\nu}} \sum_{h=1}^{H} \left(\sqrt{\frac{\operatorname{Var}_{\widehat{P}_{h,0}(s,a,b)}(\underline{V}_{h+1} - \underline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,1}(s,a,b) \vee 1}} + \frac{H\iota}{n_{h,1}(s,a,b) \vee 1} \right)$$

$$\leq c\mathbb{E}_{\mu^*,\underline{\nu}} \sum_{h=1}^{H} \left(\sqrt{\frac{cH \operatorname{Var}_{P_h(s,a,b)}(\underline{V}_{h+1} - \underline{V}_{h+1}^{\mathrm{ref}})\iota}{nd_h^{\rho}(s,a,b)}} + \frac{cH^2\iota}{nd_h^{\rho}(s,a,b)} + \frac{cH^2\iota}{nd_h^{\rho}(s,a,b)} \right)$$

$$\leq c^2\mathbb{E}_{\mu^*,\underline{\nu}} \sum_{h=1}^{H} \left(\sqrt{\frac{H\left[P_h(\underline{V}_{h+1} - \underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,a,b)\iota}{nd_h^{\rho}(s,a,b)}} + \frac{H^2\iota}{nd_h^{\rho}(s,a,b)} \right)$$

$$\leq c^2 \sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) \left(\sqrt{\frac{H\left[P_h(\underline{V}_{h+1} - \underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,a,b)\iota}{nd_m}} + \frac{H^2\iota}{nd_m} \right)$$

$$
\begin{aligned}
&\le c^2\sqrt{\sum_{h=1}^{H}\sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)}\left(\sqrt{\frac{\sum_{h=1}^{H}\sum_{(s,a,b)} H d_h^{\mu^*,\underline{\nu}}(s,a,b)\left[P_h(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,a,b)\iota}{nd_m}}+\frac{H^2\iota}{nd_m}\right) \\
&\hfill \text{(Cauchy-Schwarz Inequality)} \\
&\le c^2\sqrt{H}\sqrt{\frac{H\iota\sum_{h=1}^{H}\sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)\left[P_h(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,a,b)}{nd_m}}+\frac{c^2H^3\iota}{nd_m} \\
&\le c^2\sqrt{H\iota}\sqrt{\frac{H\iota\sum_{h=1}^{H}\sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)\left[P_h(V_{h+1}^{*}-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,a,b)}{nd_m}}+\frac{c^2H^3\iota}{nd_m} \\
&\hfill (V_{h+1}^{*}\ge \underline{V}_{h+1}\ge \underline{V}_{h+1}^{\mathrm{ref}}) \\
&= c^2\sqrt{H\iota}\sqrt{\frac{H^2\sum_{h=1}^{H}\sum_{s} d_{h+1}^{\mu^*,\underline{\nu}}(s)(V_{h+1}^{*}(s)-\underline{V}_{h+1}^{\mathrm{ref}}(s))}{nd_m}}+\frac{c^2H^3\iota}{nd_m} \\
&\le c^2\sqrt{H\iota}\sqrt{\frac{H^2 64\sqrt{\frac{H^5\iota^2}{n_{\mathrm{ref}}d_m}}}{nd_m}+\frac{c^2H^3\iota}{nd_m}} \qquad \text{(Theorem F.2)} \\
&= c^2\sqrt{\frac{192H^3\iota\sqrt{H^5\iota^2}}{(nd_m)^{3/2}}}+\frac{c^2H^3\iota}{nd_m}.
\end{aligned}
$$

□

Theorem F.5. *Suppose Assumption 3.2 holds and* $d_m = \min\{d_h^{\rho}(s,a,b) : h\in[H],(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\}$*. For any* $0<\delta<1$ *and* $n\ge \widetilde{O}(\sqrt{H^4/d_m})$*, with probability* $1-\delta$*, the output policy* $\pi=(\underline{\mu},\overline{\nu})$ *of Algorithm 2 satisfies*

$$
V_1^{*}(s_1)-V_1^{\underline{\mu},*}(s_1)\le \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right),
$$

$$
V_1^{*,\overline{\nu}}(s_1)-V_1^{*}(s_1)\le \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right).
$$

As a result, we have

$$
\mathrm{Gap}(\underline{\mu},\overline{\nu})\le \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right).
$$

Proof.

$$
\begin{aligned}
&V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1) \\
\le& 2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h)+2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,1}(s_h,a_h,b_h) && \text{(Lemma E.4)} \\
\le& \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\sqrt{V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1)}\right)+\widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right) && \text{(Lemma F.3 and Lemma F.4)} \\
\le& \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right)+\widetilde{O}\left(\frac{H^3}{nd_m}\right) && \text{(Lemma G.5)} \\
=& \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right).
\end{aligned}
$$

By the definition of NE, we have

$$V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) \le V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1) \le \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right).$$

The second argument can be proven in a similar manner. Combining two arguments together and we can derive that

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \le \widetilde{O}\left(\sqrt{\frac{H^3}{nd_m}}\right).$$

□

F.2 Turn-based Markov Game

For turn-based Markov game, there always exists a pure (deterministic) NE equilibrium strategy. As a result, we can have that μ^*, ν^*, $\underline{\mu}$, $\underline{\nu}$, $\overline{\mu}$, $\overline{\nu}$ are all pure strategy.

Theorem F.6. *Suppose Assumption 2.2 holds. For any* $0 < \delta < 1$*, with probability* $1 - \delta$*, the output policy* $\pi = (\underline{\mu}, \overline{\nu})$ *of Algorithm 1 satisfies*

$$\begin{aligned} V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) &\le 64\sqrt{\frac{C^* S H^5 \iota^2}{n}}, \\ V_1^{*,\overline{\nu}}(s_1) - V_1^*(s_1) &\le 64\sqrt{\frac{C^* S H^5 \iota^2}{n}}. \end{aligned}$$

As a result, we have

$$\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \le \widetilde{O}\left(\sqrt{\frac{C^* S H^5}{n}}\right)$$

Proof. By Lemma D.3, with probability $1 - \delta$ we have

$$\begin{aligned} & V_1^{\mu^*,*}(s_1) - V_1^{\underline{\mu},*}(s_1) \\ \le & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \underline{b}_h(\mathrm{s_h}, \mathrm{a_h}, \mathrm{b_h}) \\ = & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[4\sqrt{\frac{H^2 \iota}{n_h(s,a,b) \vee 1}} \right] \\ \le & 2\sum_{h=1}^{H} \mathbb{E}_{\mu^*,\underline{\nu}} \left[32\sqrt{\frac{H^3 \iota^2}{n d_h^{\rho}(s,a,b)}} \right] && \text{(Lemma D.1)} \\ = & 2\sum_{h=1}^{H} \sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b) \left[32\sqrt{\frac{H^3 \iota^2}{n d_h^{\rho}(s,a,b)}} \right] \\ \le & 64\sum_{h=1}^{H} \sum_{(s,a,b)} \left[\sqrt{\frac{d_h^{\mu^*,\underline{\nu}}(s,a,b) C^* H^3 \iota^2}{n}} \right] \\ = & 64\sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \left[\sqrt{\frac{d_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s)) C^* H^3 \iota^2}{n}} \right] && (\mu^*,\underline{\nu} \text{ are deterministic strategy.}) \\ \le & 64\sqrt{SH} \cdot \sqrt{\frac{\sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s)) C^* H^3 \iota^2}{n}} && \text{(Cauchy-Schwarz Inequality)} \\ = & 64\sqrt{\frac{C^* S H^5 \iota^2}{n}}. \end{aligned}$$

□

Theorem F.7. *Suppose Assumption 2.2 holds. For any $0 < \delta < 1$ and policy μ, ν, with probability $1 - \delta$, the pessimistic value $\underline{V}_h$ of Algorithm 1 satisfies*

$$\mathbb{E}_{\mu^*,\nu}\left[V_h^*(s_h) - \underline{V}_h(s_h)\right] \le 64\sqrt{\frac{C^* S H^5 \iota^2}{n}},$$

$$\mathbb{E}_{\mu,\nu^*}\left[\overline{V}_h(s_h) - V_h^*(s_h)\right] \le 64\sqrt{\frac{C^* S H^5 \iota^2}{n}},$$

where s_h is sampled from the trajectory following the strategy in the expectation at timestep h.

Proof. By Lemma D.3, under good event $\mathcal{G}$ for all state s we have

$$\begin{aligned}&V_h^*(s) - V_h^{\mu,*}(s)\\ \le& 2\sum_{t=h}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[\underline{b}_h(s_t,a_t,b_t)|s_h = s\right]\end{aligned}$$

We define $\nu' = (\nu_1, \cdots, \nu_{h-1}, \underline{\nu}_h, \cdots, \underline{\nu}_H)$. Then we have

$$\begin{aligned}\mathbb{E}_{\mu^*,\nu}\left[V_h^*(s_h) - \underline{V}_h(s_h)\right] \le& \mathbb{E}_{\mu^*,\nu}\left[2\sum_{t=h}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[\underline{b}_h(s_t,a_t,b_t)|s_h = s\right]|s\right]\\ =& 2\sum_{t=h}^{H}\mathbb{E}_{\mu^*,\nu'}\left[\underline{b}_h(s_t,a_t,b_t)\right].\end{aligned}$$

Then following the proof of Theorem F.6, we can prove the argument.

□

Lemma F.8. *For $n \ge \widetilde{O}(\sqrt{C^* S H^3})$, we have*

$$\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h) \le \widetilde{O}\left(\sqrt{\frac{C^* S H^3}{n}}\sqrt{V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1)}\right) + \widetilde{O}\left(\sqrt{\frac{C^* S H^3}{n}}\right).$$

Proof.

$$\begin{aligned}&\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h)\\ =& c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,0}(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,0}(s,a,b)\vee 1}} + \frac{H\iota}{n_{h,0}(s,a,b)\vee 1}\right)\\ \le& c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{c\,\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n d_h^{\rho}(s,a,b)}} + \frac{cH\iota}{n d_h^{\rho}(s,a,b)} + \frac{cH\iota}{n d_h^{\rho}(s,a,b)}\right)\\ =& c^2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s))\left(\sqrt{\frac{\mathrm{Var}_{P_h(s,\mu^*(s),\underline{\nu}(s))}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n d_h^{\rho}(s,\mu^*(s),\underline{\nu}(s))}} + \frac{H\iota}{n d_h^{\rho}(s,\mu^*(s),\underline{\nu}(s))}\right)\\ \le& c^2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left(\sqrt{\frac{C^* d_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s))\,\mathrm{Var}_{P_h(s,\mu^*(s),\underline{\nu}(s))}(\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n}} + \frac{C^* H\iota}{n}\right)\end{aligned}$$

(μ^*, $\underline{\nu}$ are deterministic strategies.)

$$\le c^2\sqrt{SH}\cdot\sqrt{\frac{C^*\iota\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s))\,\mathrm{Var}_{P_h(s,\mu^*(s),\underline{\nu}(s))}(\underline{V}_{h+1}^{\mathrm{ref}})}{n}} + \frac{c^2 S C^* H\iota}{n}$$

$$\leq c^2\sqrt{C^*SH\iota}\cdot\sqrt{\frac{\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}^{\mathrm{ref}})\right]}{n}+\frac{c^2SC^*H\iota}{n}}$$

$$\leq c^2\sqrt{C^*SH\iota}\cdot\sqrt{\frac{\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[\mathrm{Var}_{P_h(s,a,b)}(V_{h+1}^{\mu^*,\underline{\nu}})+2H[P_h(V_{h+1}^{\mu^*,\underline{\nu}}-\underline{V}_{h+1}^{\mathrm{ref}})](s,a,b)\right]}{n}+\frac{c^2SC^*H\iota}{n}}$$

(Lemma G.4)

$$\leq c^2\sqrt{C^*SH\iota}\cdot\sqrt{\frac{H^2+2H\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[V_{h+1}^{\mu^*,\underline{\nu}}(s_{h+1})-\underline{V}_{h+1}^{\mathrm{ref}}(s_{h+1})\right]}{n}+\frac{c^2SC^*H\iota}{n}}$$

(Lemma E.5)

$$=c^2\sqrt{C^*SH\iota}\cdot\sqrt{\frac{H^2+2H\sum_{h=1}^{H}\mathbb{E}_{\mu^*,\underline{\nu}}\left[V_{h+1}^{\mu^*,\underline{\nu}}(s_{h+1})-V_{h+1}^{*}(s_{h+1})+V_{h+1}^{*}(s_{h+1})-\underline{V}_{h+1}^{\mathrm{ref}}(s_{h+1})\right]}{n}+\frac{c^2SC^*H\iota}{n}}$$

$$\leq c^2\sqrt{C^*SH\iota}\cdot\sqrt{\frac{H^2+2H^2(V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1))+128H\sqrt{\frac{C^*SH^5\iota^2}{n_{\mathrm{ref}}}}}{n}+\frac{c^2SC^*H\iota}{n}}$$

(Lemma E.6 and Theorem F.7)

$$\leq\frac{c^2\sqrt{C^*SH^3\iota}}{\sqrt{n}}+\frac{c^2\sqrt{384C^*SH^2\iota\sqrt{C^*SH^5\iota^2}}}{n^{3/4}}+\frac{c\sqrt{2C^*SH^3\iota}}{\sqrt{n}}\sqrt{V_1^{\mu^*,\underline{\nu}}(s_1)-\underline{V}_1(s_1)}+\frac{c^2SC^*H\iota}{n}.$$

□

Lemma F.9. *For $n\geq\widetilde{O}(\sqrt{C^*SABH^4})$, we have*

$$\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,1}(s_h,a_h,b_h)\leq\widetilde{O}\left(\sqrt{\frac{C^*SABH^3}{n}}\right).$$

Proof.

$$\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,1}(s_h,a_h,b_h)$$

$$=c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{\mathrm{Var}_{\widehat{P}_{h,0}(s,a,b)}(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})\iota}{n_{h,1}(s,a,b)\vee 1}}+\frac{H\iota}{n_{h,1}(s,a,b)\vee 1}\right)$$

$$\leq c\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{cH\,\mathrm{Var}_{P_h(s,a,b)}(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})\iota}{nd_h^{\rho}(s,a,b)}}+\frac{cH^2\iota}{nd_h^{\rho}(s,a,b)}+\frac{cH^2\iota}{nd_h^{\rho}(s,a,b)}\right)$$

$$\leq c^2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\left(\sqrt{\frac{H\left[P_h(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,a,b)\iota}{nd_h^{\rho}(s,a,b)}}+\frac{H^2\iota}{nd_h^{\rho}(s,a,b)}\right)$$

$$=c^2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}d_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s))\left(\sqrt{\frac{H\left[P_h(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,\mu^*(s),\underline{\nu}(s))\iota}{nd_h^{\rho}(s,\mu^*(s),\underline{\nu}(s))}}+\frac{H^2\iota}{nd_h^{\rho}(s,\mu^*(s),\underline{\nu}(s))}\right)$$

$$\leq c^2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left(\sqrt{\frac{C^*Hd_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s))\left[P_h(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,\mu^*(s),\underline{\nu}(s))\iota}{n_1}}+\frac{H^2C^*\iota}{n_1}\right)$$

(Cauchy-Schwarz Inequality)

$$
\begin{aligned}
\leq & c^2\sqrt{SH\iota}\sqrt{\frac{C^*H\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\mu^*,\underline{\nu}}(s,\mu^*(s),\underline{\nu}(s))\left[P_h(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,\mu^*(s),\underline{\nu}(s))}{n}} + \frac{c^2SH^3C^*\iota}{n} \\
\leq & c^2\sqrt{SH\iota}\sqrt{\frac{C^*H\iota\sum_{h=1}^{H}\sum_{(s,a,b)} d_h^{\mu^*,\underline{\nu}}(s,a,b)\left[P_h(V_{h+1}^*-\underline{V}_{h+1}^{\mathrm{ref}})^2\right](s,a,b)}{n}} + \frac{c^2C^*SH^3\iota}{n} \\
& \qquad (V_{h+1}^* \geq \underline{V}_{h+1} \geq \underline{V}_{h+1}^{\mathrm{ref}}) \\
= & c^2\sqrt{SH\iota}\sqrt{\frac{H^2C^*\sum_{h=1}^{H}\sum_s d_{h+1}^{\mu^*,\underline{\nu}}(s)(V_{h+1}^*(s)-\underline{V}_{h+1}^{\mathrm{ref}}(s))}{n}} + \frac{c^2C^*SH^3\iota}{n} \\
\leq & c^2\sqrt{SH\iota}\sqrt{\frac{H^2C^*64\sqrt{\frac{C^*SH^5\iota^2}{n_{\mathrm{ref}}}}}{n}} + \frac{c^2SH^3C^*\iota}{n} \qquad \text{(Theorem F.7)} \\
= & c^2\sqrt{\frac{192C^*SH^3\iota\sqrt{C^*SH^5\iota^2}}{n^{3/2}}} + \frac{c^2C^*SH^3\iota}{n}.
\end{aligned}
$$

□

Theorem F.10. *Suppose Assumption 2.2 holds for a turn-based Markov game. For any $0 < \delta < 1$, with probability $1-\delta$, the output policy $\pi = (\underline{\mu}, \overline{\nu})$ of Algorithm 1 satisfies*

$$
V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) \leq \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\right),
$$

$$
V_1^{*,\overline{\nu}}(s_1) - V_1^*(s_1) \leq \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\right).
$$

As a result, we have

$$
\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \leq \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\right).
$$

Proof.

$$
\begin{aligned}
& V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1) \\
\leq & 2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,0}(s_h,a_h,b_h) + 2\mathbb{E}_{\mu^*,\underline{\nu}}\sum_{h=1}^{H}\underline{b}_{h,1}(s_h,a_h,b_h) \qquad \text{(Lemma E.4)} \\
\leq & \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\sqrt{V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1)}\right) + \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\right) \qquad \text{(Lemma F.8 and Lemma F.9)} \\
\leq & \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\right) + \widetilde{O}\left(\frac{C^*SH^3}{n}\right) \qquad \text{(Lemma G.5)} \\
= & \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\right).
\end{aligned}
$$

By the definition of NE, we have

$$
V_1^*(s_1) - V_1^{\underline{\mu},*}(s_1) \leq V_1^{\mu^*,\underline{\nu}}(s_1) - \underline{V}_1(s_1) \leq \widetilde{O}\left(\sqrt{\frac{C^*SABH^3}{n}}\right).
$$

The second argument can be proven in a similar manner. Combining these two arguments and we can derive that

$$
\mathrm{Gap}(\underline{\mu}, \overline{\nu}) \leq \widetilde{O}\left(\sqrt{\frac{C^*SH^3}{n}}\right).
$$

□

G Auxiliary Lemmas

Lemma G.1. *(Multiplicative Chernoff bound). Let X be a binomial random variable with parameter p, n. For any $1 \geq \theta > 0$, we have that*

$$\mathbb{P}[(1-\theta)pn < X < (1+\theta)pn] < 2e^{-\frac{\theta^2 pn}{2}}$$

Lemma G.2. *For all $(\mathrm{s_h}, \mathrm{a_h}, \mathrm{b_h}) \in \mathcal{K}_h$ and any $\|V\|_\infty \leq H$, with probability $1-\delta$ we have*

$$\sqrt{\operatorname*{Var}_{\widehat{P}^\dagger_{\mathrm{s_h},\mathrm{a_h},\mathrm{b_h}}}(V)} \leq \sqrt{\operatorname*{Var}_{P^\dagger_{\mathrm{s_h},\mathrm{a_h},\mathrm{b_h}}}(V)} + cH\sqrt{\frac{\iota}{nd^\mu_h(\mathrm{s_h}, \mathrm{a_h}, \mathrm{b_h})}}.$$

Proof. The is a direct application of Lemma G.3 with a union bound. □

Lemma G.3. *(Empirical Berstein Inequality (Maurer & Pontil, 2009)) Let $n \geq 2$ and $V \in \mathbb{R}^S$ be any functions with $\|V\|_\infty \leq H$, P be any S-dimensional distribution and $\widehat{P}$ be its empirical version using n samples. Then with probability $1-\delta$,*

$$\left|\sqrt{\operatorname*{Var}_{\widehat{P}}(V)} - \sqrt{\frac{n-1}{n}\operatorname*{Var}_{P}(V)}\right| \leq 2H\sqrt{\frac{\log(2/\delta)}{n-1}}.$$

Lemma G.4. *For $0 \leq V \leq V' \leq H$, we have*

$$\operatorname*{Var}_{P_h(s,a,b)}(V) \leq \operatorname*{Var}_{P_h(s,a,b)}(V') + 2H[P_h(V'-V)](s,a,b).$$

Proof.

$$\begin{aligned}
&\operatorname*{Var}_{P_h(s,a,b)}(V) - \operatorname*{Var}_{P_h(s,a,b)}(V') \\
\leq &\left[P_h(V)^2 - (P_hV)^2 - P_h(V')^2 + (P_hV')^2\right](s,a,b) \\
= &\left[P_h(V+V')(V-V') + [P_h(V'-V)][P_h(v'+v)]\right](s,a,b) \\
\leq &2H[P_h(V'-V)](s,a,b).
\end{aligned}$$

□

Lemma G.5. *If $x \leq a\sqrt{x} + b$ for $a, b > 0$, then we have*

$$x \leq 2a^2 + 2b.$$

Proof. We have

$$(\sqrt{x} - \frac{a}{2})^2 \leq b + \frac{a^2}{4}.$$

If $\sqrt{x} < \frac{a}{2}$, the argument holds directly. Otherwise we have

$$\sqrt{x} - \frac{a}{2} \leq \sqrt{b + \frac{a^2}{4}} \leq \sqrt{b} + \frac{a}{2}.$$

So we have $\sqrt{x} \leq \sqrt{b} + a$, which implies $x \leq 2(a^2 + b)$. □