Improving Sample Efficiency of Model-Free Algorithms for Zero-Sum Markov Games

Songtao Feng [1] **Ming Yin** [2] **Yu-Xiang Wang** [3] **Jing Yang** [4] **Yingbin Liang** [5]

Abstract

The problem of two-player zero-sum Markov games has recently attracted increasing interests in theoretical studies of multi-agent reinforcement learning (RL). In particular, for finite-horizon episodic Markov decision processes (MDPs), it has been shown that model-based algorithms can find an ϵ-optimal Nash Equilibrium (NE) with the sample complexity of $O(H^3SAB/\epsilon^2)$, which is optimal in the dependence of the horizon H and the number of states S (where A and B denote the number of actions of the two players, respectively). However, none of the existing model-free algorithms can achieve such an optimality. In this work, we propose a model-free stage-based algorithm and show that it achieves the same sample complexity as the best model-based algorithm, and hence for the first time demonstrate that model-free algorithms can enjoy the same optimality in the H dependence as model-based algorithms. The main improvement of the dependency on H arises by leveraging the popular variance reduction technique based on the reference-advantage decomposition previously used only for single-agent RL. However, such a technique relies on a critical monotonicity property of the value function, which does not hold in Markov games due to the update of the policy via the coarse correlated equilibrium (CCE) oracle. Thus, to extend such a technique to Markov games, our algorithm features a key novel design of updating the reference value functions as the pair of optimistic and pessimistic value functions whose value difference is the smallest in the history in order to achieve the desired improvement in the sample efficiency.

[1]The University of Florida [2]Princeton University [3]The University of California, Santa Barbara [4]The Pennsylvania State University [5]The Ohio State University. Correspondence to: Songtao Feng <sfeng1@ufl.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Multi-agent reinforcement learning (MARL) commonly refers to the sequential decision making framework, in which more than one agent learn to make decisions in an unknown shared environment to maximize their cumulative rewards. MARL has achieved great success in a variety of practical applications, including the game of GO [Silver et al., 2016; 2017], real-time strategy games involving team play [Vinyals et al., 2019], autonomous driving [Shalev-Shwartz et al., 2016], and behavior learning in complex social scenarios [Baker et al., 2020]. Despite the great empirical success, one major bottleneck for many RL algorithms is that they require enormous samples. For example, in many practical MARL scenarios, a large number of samples are often required to achieve human-like performance due to the necessity of exploration. It is thus important to understand how to design sample-efficient algorithms.

As a prevalent approach to the MARL, model-based methods use the existing visitation data to estimate the model, run a planning algorithm on the estimated model to obtain the policy, and execute the policy in the environment. In two-player zero-sum Markov games, an extensive series of studies [Bai & Jin, 2020a; Zhang et al., 2020a; Liu et al., 2021] have shown that *model-based* algorithms are provably efficient in MARL, and can achieve minimax-optimal sample complexity $O(H^3SAB/\epsilon^2)$ except for the term AB [Zhang et al., 2020a; Liu et al., 2021], where H denotes the horizon, S denotes the number of states, and A and B denote the numbers of actions of the two players, respectively. On the other hand, *model-free* methods directly estimate the (action-)value functions at the equilibrium policies instead of estimating the model. However, none of the existing *model-free* algorithms can achieve the aforementioned optimality (attained by model-based algorithms) [Bai et al., 2020; Mao & Başar, 2021; Song et al., 2022; Jin et al., 2022a; Mao et al., 2022]. Specifically, the number of episodes required for model-free algorithms scales sub-optimally in step H, which naturally motivates the following open question:

Can we design model-free algorithms with the optimal sample dependence on the time horizon for learning two-player zero-sum Markov games?

In this paper, we give an affirmative answer to the above question. We highlight our main contributions as follows.

Algorithm design. We design a new model-free algorithm of Q-learning with **min-gap** based reference-advantage decomposition. In particular, we extend the reference-advantage decomposition technique [Zhang et al., 2020b] proposed for single-agent RL to zero-sum Markov games with the following key novel design. Unlike the single-agent scenario, the optimistic (or pessimistic) value function in Markov games does not necessarily preserve the monotone property due to the nature of the CCE oracle. In order to obtain the "best" optimistic and pessimistic value function pair, we update the reference value functions as the pair of optimistic and pessimistic value functions whose value difference is the smallest (i.e., with the minimal gap) in the history. Moreover, our algorithm relies on the stage-based approach, which simplifies the algorithm design and subsequent analysis.

Sample complexity bound. We show that our algorithm provably finds an ϵ-optimal Nash equilibrium for the two-player zero-sum Markov game in $\widetilde{O}(H^3SAB/\epsilon^2)$ episodes, which improves upon the sample complexity of all existing model-free algorithms for zero-sum Markov game. Further, comparison to the existing lower bound shows that it is minimax-optimal on the dependence of H, S and ϵ. This is the first result that establishes such optimality for model-free algorithms, although model-based algorithms have been shown to achieve such optimality in the past [Liu et al., 2021].

Technical analysis. We establish a few new properties on the cumulative occurrence of the large V-gap and the cumulative bonus term to enable the upper-bounding of several new error terms arising due to the incorporation of the new min-gap based reference-advantage decomposition technique. These properties have not been established for the single-agent RL with such a technique, because our properties are established for policies generated by the CCE oracle in zero-sum Markov games. Further, the analysis of both the optimistic and pessimistic accumulative bonus terms requires a more refined analysis compared to their counterparts in single-agent RL [Zhang et al., 2020b].

1.1. Related Work

Markov games. The Markov game, also known as the stochastic game, was first proposed in [Shapley, 1953] to model the multi-agent RL. Early attempts to find the Nash equilibra of Markov games include [Littman, 1994; Hu & Wellman, 2003; Hansen et al., 2013; Wei et al., 2020]. However, they often relied on strong assumptions such as known transition matrix and reward, or focused on the asymptotic setting. Thus, these results do not apply to the non-asymptotic setting where the transition and reward are unknown and only limited data is available.

There is a line of works focusing on non-asymptotic guarantees with certain reachability assumptions. A popular approach is to assume access to simulators, which enables the agent to sample transition and reward directly for any state-action pair [Jia et al., 2019; Sidford et al., 2020; Zhang et al., 2020a; Li et al., 2022]. Alternatively, [Wei et al., 2017] studied the Markov game under the assumption that one player can always reach all states by playing certain policy no matter what strategy the other player sticks to.

Two-player zero-sum games. [Bai & Jin, 2020a; Xie et al., 2020] initialized the study of non-asymptotic guarantee for two-player zero-sum Markov games without reachability assumptions. [Bai & Jin, 2020a] proposed a model-based algorithm for tabular Markov game while [Xie et al., 2020] considered linear function approximation in game and adopted a model-free approach. [Liu et al., 2021] proposed a model-based algorithm which achieves the minimax-optimal samples complexity $O(H^3SAB/\epsilon)$ except for the AB term. For the discounted setting and having access to a generative model, [Zhang et al., 2020a] developed a model-based algorithm that achieves the minimax-optimal sample complexity except for the AB term. Then, model-free Nash Q-learning and Nash V-learning were proposed in [Bai et al., 2020] for two-player zero-sum game to achieve optimal dependence on actions (i.e., $(A+B)$ instead of AB). Further, [Chen et al., 2022; Huang et al., 2022] studied the two-player zero-sum Markov game under linear and general function approximation.

Multi-player general-sum games. [Liu et al., 2021] developed model-free algorithm in episodic setting, which suffers from the curse of multi-agent. To alleviate this issue, [Mao & Başar, 2021; Song et al., 2022; Jin et al., 2022a; Mao et al., 2022] proposed V-learning algorithm, coupled with the adversarial bandit subroutine, to break the curse of multi-agent. [Mao & Başar, 2021] considered learning an ϵ-optimal CCE and used V-learning with stabilized online mirror descent as the adversarial bandit subroutine. Both [Song et al., 2022; Jin et al., 2022a] utilized the weighted follow the regularized leader (FTRL) algorithm as the adversarial subroutine, and considered ϵ-optimal CCE and ϵ-optimal correalted equilibrium (CE). The work [Mao et al., 2022] featured the standard uniform weighted FTRL and staged-based design, both of which simplifies the algorithm design and the corresponding analysis. While the V-learning algorithms generate non-Markov, history dependent policies, [Daskalakis et al., 2022; Wang et al., 2023] learned an approximate CCEs that is guaranteed to be Markov.

Markov games with function approximation. Recently, a few works considered learning in Markov games with linear function approximation [Xie et al., 2020; Chen et al., 2022] and general function approximation [Jin et al., 2022b;

Huang et al., 2022; Zhan et al., 2023; Xiong et al., 2022; Chen et al., 2022; Ni et al., 2023]. While all of the previous works require centralized function classes and inevitably suffer from the curse of multi-agency, [Cui et al., 2023; Wang et al., 2023] proposed decentralized MARL algorithms to resolve the issue under linear and general function approximation.

Single-agent RL. Broadly speaking, our work is also related to single-agent RL [Auer et al., 2008; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Zhang et al., 2020b]. As a special case of Markov games, only one agent interacts with the environment in single-agent RL. For tabular episodic setting, the minimax-optimal sample complexity is $\widetilde{O}(H^3SA/\epsilon^2)$, achieved by a model-based algorithm in [Azar et al., 2017] and a model-free algorithm in [Zhang et al., 2020b]. Technically, the reference-advantage decomposition used in our algorithm is similar to that of [Zhang et al., 2020b], as both employ variance reduction techniques for faster convergence. However, our approaches differ significantly, particularly in the way of handling the interplay between the CCE oracle and the reference-advantage decomposition in the context of two-player zero-sum Markov game.

2. Preliminaries

Zero-sum Markov Game. We consider the tabular episodic two-player zero-sum Markov game $\mathrm{MG}(H, \mathcal{S}, \mathcal{A}, \mathcal{B}, P, r)$, where H is the number of steps in each episode, $\mathcal{S}$ is the set of states with $|\mathcal{S}| = S$, $(\mathcal{A}, \mathcal{B})$ are the sets of actions of the max-player and the min-player respectively with $|\mathcal{A}| = A$ and $|\mathcal{B}| = B$, $P = \{P_h\}_{h\in[H]}$ is the collection of the transition matrices with $P_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \mapsto \mathcal{S}$, $r = \{r_h\}_{h\in[H]}$ is the collection of deterministic reward functions with $r_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \mapsto [0, 1]$. Here the reward represents both the gain of the max-player and the loss of the min-player. We assume each episode starts with a fixed initial state s_1.

Suppose the max-player and the min-player interact with the environment sequentially captured by the two-player zero-sum Markov game $\mathrm{MG}(H, \mathcal{S}, \mathcal{A}, \mathcal{B}, P, r)$. At each step $h \in [H]$, both players observe the state $s_h \in \mathcal{S}$, take their actions $a_h \in \mathcal{A}$ and $b_h \in \mathcal{B}$ simultaneously, receive the reward $r_h(s_h, a_h, b_h)$, and then the Markov game evolves into the next state with probability $s_{h+1} \sim P_h(\cdot|s_h, a_h, b_h)$. The episode ends when s_{H+1} is reached.

Markov policy, value function. A Markov policy μ of the max-player is the collection of the functions $\{\mu_h : \mathcal{S} \mapsto \Delta_{\mathcal{A}}\}_{h\in[H]}$, each of which maps from a state to a distribution over actions. Similarly, a policy ν of the min-player is the collection of functions $\{\nu_h : \mathcal{S} \mapsto \Delta_{\mathcal{B}}\}_{h\in[H]}$. We use $\mu_h(a|s)$ and $\nu_h(b|s)$ to denote the probability of taking actions a and b given the state s under the Markov policies μ and ν at step h, respectively.

Given a max-player policy μ, a min-player policy ν, and a state s at step h, the value function is defined as

$$V_h^{\mu,\nu}(s) = \mathop{\mathbb{E}}_{(s_{h'},a_{h'},b_{h'})\sim(\mu,\nu)}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})\middle| s_h = s\right].$$

For a given $(s, a, b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ under a max-player policy μ and a min-player policy ν at step h, we define

$$Q_h^{\mu,\nu}(s, a, b) = \mathop{\mathbb{E}}_{(s_{h'},a_{h'},b_{h'})\sim(\mu,\nu)}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})\middle| s_h = s, a_h = a, b_h = b\right].$$

For ease of exposition, we define $(P_h f)(s, a, b) = \mathbb{E}_{s'\sim P_h(\cdot|s,a,b)}[f(s')]$ for any function $f : \mathcal{S} \mapsto \mathbb{R}$, and $(\mathbb{D}_\pi g)(s) = \mathbb{E}_{(a,b)\sim\pi(\cdot,\cdot|s)}[g(s, a, b)]$ for any function $g : \mathcal{S}\times\mathcal{A}\times\mathcal{B}$. Then, the following Bellman equations hold for all $(s, a, b, h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$:

$$\begin{aligned}
&Q_h^{\mu,\nu}(s, a, b) = (r_h + P_h V_{h+1}^{\mu,\nu})(s, a, b),\\
&V_h^{\mu,\nu}(s) = (\mathbb{D}_{\mu_h\times\nu_h} Q_h^{\mu,\nu})(s),\\
&V_{H+1}^{\mu,\nu}(s) = 0.
\end{aligned}$$

Best response, Nash equilibrium (NE). For any Markov policy μ of the max-player, there exists a *best response* of the min-player, which is a policy $\nu^\dagger(\mu)$ satisfying $V_h^{\mu,\nu^\dagger(\mu)}(s) = \inf_\nu V_h^{\mu,\nu}$ for any $(s, h) \times \mathcal{S}\times[H]$. We denote $V_h^{\mu,\dagger} = V_h^{\mu,\nu^\dagger(\mu)}$. Similarly, the best response of the max-player with respect to the Markov policy ν of the min-player is a policy $\mu^\dagger(\nu)$ satisfying $V_h^{\mu^\dagger(\nu),\nu}(s) = \sup_\mu V_h^{\mu,\nu}$ for any $(s, h) \times \mathcal{S}\times[H]$, and we use $V_h^{\dagger,\nu}$ to denote $V_h^{\mu^\dagger(\nu),\nu}$. Further, there exists Markov policies μ^*, ν^*, which are optimal against the best responses of the other player [Filar & Vrieze, 1997], i.e.,

$$\begin{aligned}
V_h^{\mu^*,\dagger}(s) &= \sup_\mu V_h^{\mu,\dagger}(s),\\
V_h^{\dagger,\nu^*}(s) &= \inf_\nu V_h^{\dagger,\nu},
\end{aligned}$$

for all $(s, h) \in \mathcal{S}\times[H]$. We call the strategies (μ^*, ν^*) the *Nash equilibrium* of a Markov game, if they satisfy the following minimax equation

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

Learning objective. We consider the Nash equilibrium of Markov games. We measure the sub-optimality of any pair of general policies (μ, ν) using the following gap between

their performance and the performance of the optimal strategy (i.e., Nash equilibrium) when playing against the best responses respectively:

$$V_1^{\dagger,\nu}(s_1) - V_1^{\mu,\dagger}(s_1) = \left(V_1^{\dagger,\nu}(s_1) - V_1^*(s_1)\right) + \left(V_1^*(s_1) - V_1^{\mu,\dagger}(s_1)\right).$$

Definition 2.1 (ϵ-optimal Nash equilibrium (NE))**.** A pair of general policies (μ, ν) is an ϵ-optimal Nash equilibrium if $V_1^{\dagger,\nu}(s_1) - V_1^{\mu,\dagger}(s_1) \le \epsilon$.

Our goal is to design algorithms for two-player zero-sum Markov games that can find an ϵ-optimal NE using a number episodes that is small in its dependency on S, A, B, H as well as $1/\epsilon$.

3. Algorithm Design

In this section, we propose an algorithm called Q-learning with min-gap based reference-advantage decomposition (Algorithm 1), for learning ϵ-optimal Nash Equilibrium in two-player zero-sum Markov games. Our algorithm builds upon the Nash Q-learning framework [Bai et al., 2020] for two-player zero-sum Markov game but incorporates a novel **min-gap** based reference-advantage decomposition technique and stage-based update design, which were originally proposed to achieve optimal performance in model-free single-agent RL. We start by reviewing the algorithm with reference-advantage decomposition in single agent RL [Zhang et al., 2020b].

Reference-advantage decomposition in single-agent RL. In single-agent RL, we greedily select and action to maximize the action value function $\overline{Q}_h(s,a)$ to obtain the optimistic value function $\overline{V}_h(s) = \max_a \overline{Q}_h(s,a)$, and the action-value function update follows $\overline{Q}_h(s,a) \leftarrow \min\{\overline{Q}_h^{(1)}(s,a), \overline{Q}_h^{(2)}(s,a), \overline{Q}_h(s,a)\}$, where $\overline{Q}_h^{(1)}, \overline{Q}_h^{(2)}$ represent the standard update rule and the advantage-based update rule

$$\overline{Q}_h^{(1)} \leftarrow r_h(s,a) + \widehat{P_h \overline{V}_{h+1}}(s,a) + \mathrm{bonus}_1,$$
$$\overline{Q}_h^{(2)} \leftarrow r_h(s,a) + \widehat{P_h \overline{V}_{h+1}^{\mathrm{ref}}}(s,a) + \widehat{P_h(\overline{V}_{h+1} - \overline{V}_{h+1}^{\mathrm{ref}})}(s,a) + \mathrm{bonus}_2.$$

In standard update rule, one major drawback is that the early samples collected for estimating $\overline{V}_{h+1}$ at that moment deviates from the true value of $\overline{V}_{h+1}$, and we have to only use the latest samples to estimate $\widehat{P_h \overline{V}_{h+1}}(s,a)$ in order not to ruin the whole estimate, which leads to the suboptimal sample complexity of such an algorithm. To achieve the optimal sample complexity, reference-advantage decomposition was introduced. At high level, we first learn an accurate estimation $\overline{V}_h^{\mathrm{ref}}$ of the optimal value function V_h^* satisfying $V_h^*(s) \le V_h^{\mathrm{ref}}(s) \le V_h^*(s) + \beta$, where the accuracy is controlled by parameter β independent of the number of episodes K. For the second term, since $\overline{V}_{h+1}^{\mathrm{ref}}$ is almost fixed, we are able to conduct the estimate using all collected samples. For the third term, we still have to only use the latest samples to limit the deviation error. Thanks to the reference-advantage decomposition, and since $\overline{V}_{h+1}$ is learned based on $\overline{V}_{h+1}^{\mathrm{ref}}$, and $\overline{V}_{h+1}^{\mathrm{ref}}$ is already an accurate estimate of V_{h+1}^*, it turns out that estimating $\overline{V}_{h+1} - \overline{V}_{h+1}^{\mathrm{ref}}$ instead of directly estimating $\overline{V}_{h+1}$ offsets the weakness of using the latest samples.

In single-agent RL, one key design to facilitate the reference-advantage decomposition is to ensure that the action-value function $\overline{Q}_h(s,a)$ is non-increasing. Observe that the optimistic value function $\overline{V}_h(s)$ preserves the monotonic structure as long as the optimistic action-value function $\overline{Q}_h(s)$ is non-increasing, since $\overline{V}_h^{k+1}(s) = \max_a \overline{Q}_h^{k+1}(s,a) \le \max_a \overline{Q}_h^k(s,a) = \overline{V}_h^k(s)$. When enough samples are collected, the reference value $\overline{V}^{\mathrm{ref}}$ is then updated as the latest optimistic value function, which we remark is also the smallest optimistic value function in the up-to-date learning history.

Min-gap[1] based reference-advantage decomposition. In the two-player zero-sum game, we keep track of both the optimistic and the pessimistic action-value functions, and update the value functions using the CCE oracle at the end of each stage. Unlike the single-agent scenario, the optimistic (or pessimistic) value function does not necessarily preserve the monotone property even if the optimistic (or pessimistic) action-value function is non-increasing (or non-decreasing) due to the nature of the CCE oracle. In order to obtain the "best" optimistic and pessimistic value function pair, we come up with the key novel "**min-gap**" design where we update the reference value functions as the pair of optimistic and pessimistic value functions whose value difference is the smallest in the history (line 12-15). Formally, we define the min-gap $\Delta(s,h)$ for a state s at step h to keep track of the smallest value difference between optimistic and pessimistic value functions in the history, and the corresponding pair of value functions are recorded (line 12-13). When enough samples are collected (line 14-15), the pair of reference value functions is then set to be the pair of optimistic and pessimistic value functions whose value difference is the smallest in the history.

Now we introduce reference-advantage decomposition to the two-player zero-sum game. For ease of exposition, we use bonus_i to represent different exploration bonus, which is specified in line 9-11 of Algorithm 3. In standard update

[1] We remark that min-gap has nothing to do with the notion of gap in gap-dependent RL.

Algorithm 1 Q-learning with min-gap based reference-advantage decomposition (Algorithm 3 sketch)

1: Set accumulators and (action)-value functions properly, and initialize the gap $\Delta(s,h)=H$.
2: **for** episodes $k \leftarrow 1,2,\ldots,K$ **do**
3: **for** $h \leftarrow 1,2,\ldots,H$ **do**
4: Take action $(a_h,b_h) \leftarrow \pi_h(s_h)$
5: Receive $r_h(s_h,a_h,b_h)$, and observe s_{h+1}.
6: Update accumulators.
7: **if** $n \in \mathcal{L}$ **then**
8: $\overline{Q}_h(s_h,a_h,b_h) \leftarrow \min\{\overline{Q}_h^{(1)}(s_h,a_h,b_h),$
9: $\overline{Q}_h^{(2)}(s_h,a_h,b_h), \overline{Q}_h(s_h,a_h,b_h)\}.$
10: $\underline{Q}_h(s_h,a_h,b_h) \leftarrow \max\{\underline{Q}_h^{(1)}(s_h,a_h,b_h),$
11: $\underline{Q}_h^{(2)}(s_h,a_h,b_h), \underline{Q}_h(s_h,a_h,b_h)\}.$
12: $\pi_h(s_h) \leftarrow \mathrm{CCE}(\overline{Q}(s_h,\cdot,\cdot), \underline{Q}_h(s_h,\cdot,\cdot))$.
13: $\overline{V}_h(s_h) \leftarrow \mathbb{E}_{(a,b)\sim\pi_h(s_h)}\overline{Q}_h(s_h,a,b)$.
14: $\underline{V}_h(s_h) \leftarrow \mathbb{E}_{(a,b)\sim\pi_h(s_h)}\underline{Q}_h(s_h,a,b)$.
15: Reset all intra-stage accumulators to 0.
16: **if** $\overline{V}_h(s_h)-\underline{V}_h(s_h) < \Delta(s,h)$ **then**
17: $\Delta(s,h)=\overline{V}_h(s_h)-\underline{V}_h(s_h)$.
18: $\widetilde{\overline{V}}_h(s_h)=\overline{V}_h(s_h)$.
19: $\widetilde{\underline{V}}_h(s_h)=\underline{V}_h(s_h)$.
20: **end if**
21: **end if**
22: **if** $\sum_{a,b} N_h(s_h,a,b)=N_0$ **then**
23: $\overline{V}_h^{\mathrm{ref}}(s_h) \leftarrow \widetilde{\overline{V}}_h(s_h)$.
24: $\underline{V}_h^{\mathrm{ref}}(s_h) \leftarrow \widetilde{\underline{V}}_h(s_h)$.
25: **end if**
26: **end for**
27: **end for**

rule, we have

$$\overline{Q}_h^{(1)}(s,a,b) \leftarrow r_h(s,a,b)+\widehat{P_h\overline{V}_{h+1}}(s,a,b)+\mathrm{bonus}_3, \quad (1)$$

$$\underline{Q}_h^{(1)}(s,a,b) \leftarrow r_h(s,a,b)+\widehat{P_h\underline{V}_{h+1}}(s,a,b)+\mathrm{bonus}_4, \quad (2)$$

where $\widehat{P_h\overline{V}_{h+1}}, \widehat{P_h\underline{V}_{h+1}}$ are the empirical estimate of $P_h\overline{V}_{h+1}, P_h\underline{V}_{h+1}$. Similar to the single-agent RL, the standard update rule suffers from the large deviation between $\overline{V}_{h+1}$ learned by the early samples and the value of Nash equilibrium. As a result, we have to use only the samples from the last stage (i.e., the latest $O(1/H)$ fraction of samples, see stage-based update approach below) to estimate $P_h\overline{V}_{h+1}$. In order to improve the horizon dependence, we incorporate the advantage-based update rule

$$\overline{Q}_h^{(2)}(s,a,b) \leftarrow r_h(s,a,b)+\widehat{P_h\overline{V}_{h+1}^{\mathrm{ref}}}(s,a,b) + \widehat{P_h(\overline{V}_{h+1}-\overline{V}_{h+1}^{\mathrm{ref}})}(s,a,b)+\mathrm{bonus}_5, \quad (3)$$

$$\underline{Q}_h^{(2)}(s,a,b) \leftarrow r_h(s,a,b)+\widehat{P_h\underline{V}_{h+1}^{\mathrm{ref}}}(s,a,b) + \widehat{P_h(\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}})}(s,a,b)+\mathrm{bonus}_6, \quad (4)$$

where the middle terms in (3) are the empirical estimates of $P_h\overline{V}_{h+1}^{\mathrm{ref}}$ and $P_h(\overline{V}_{h+1}-\overline{V}_{h+1}^{\mathrm{ref}})$, and the middle terms in (4) are the empirical estimates of $P_h\underline{V}_{h+1}^{\mathrm{ref}}$ and $P_h(\underline{V}_{h+1}-\overline{V}_{h+1}^{\mathrm{ref}})$. We still need to use only the samples from the last stage to limit the deviation for the third terms in both (3) and (4). For ease of exposition, assume we have access to a β-optimal $\overline{V}^{\mathrm{ref}}, \underline{V}^{\mathrm{ref}}$. Thanks to the min-gap based reference-advantage decomposition, the learned $\overline{V}_{h+1}$ (or $\underline{V}_{h+1}$) is learned based on $\overline{V}_{h+1}^{\mathrm{ref}}$ (or $\underline{V}_{h+1}^{\mathrm{ref}}$), and $\overline{V}^{\mathrm{ref}}$ (or $\underline{V}^{\mathrm{ref}}$) is already an accurate estimate of V_{h+1}^*, it turns out that estimating $\overline{V}_{h+1}-\overline{V}_{h+1}^{\mathrm{ref}}$ (or $\underline{V}_{h+1}-\underline{V}_{h+1}^{\mathrm{ref}}$) instead of directly estimating $\overline{V}$ (or $\underline{V}$) offsets the weakness of using only $O(1/H)$ fraction of data. Further, since $\overline{V}^{\mathrm{ref}}, \underline{V}^{\mathrm{ref}}$ is fixed, we are able to use all samples collected to estimate the second term, without suffering any deviation. Now we remove the assumption that $\overline{V}^{\mathrm{ref}}, \underline{V}^{\mathrm{ref}}$ is fixed. Note that β is selected independently of K. Therefore, learning a β-optimal reference value function $\overline{V}^{\mathrm{ref}}, \underline{V}^{\mathrm{ref}}$ only incurs lower order terms in our final result.

Stage-based update approach. For each tuple $(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$, we divide the visitations for the tuple into consecutive stages. The length of each stage increases exponentially with a growth rate $(1+1/H)$. Specifically, we define $e_1=H$, and $e_{i+1}=\lfloor(1+1/H)e_i\rfloor$ for all $i\geq 1$, to denote the lengths of stages. Further, we also define $\mathcal{L}=\{\sum_{i=1}^j e_i | j=1,2,3,\ldots\}$ to denote the the set of ending indices of the stages. For each (s,a,b,h) tuple, we update both the optimistic and pessimistic value estimates at the end of each stage (i.e., when the total number of visitations of (s,a,b,h) lies in $\mathcal{L}$), using samples only from this single stage (line 6-15). This updating rule ensures that only the last $O(1/H)$ fraction of the collected samples are used to estimate the value estimates.

Coarse correlated equilibrium (CCE). We use the CCE oracle to update the policy (line 14). The CCE oracle was first introduced in [Xie et al., 2020] and an ϵ-optimal CCE is shown to be a $O(\epsilon)$-optimal Nash equilibrium in two-player zero-sum Markov games [Xie et al., 2020]. For any pair of matrices $\overline{Q}, \underline{Q} \in [0,H]^{A\times B}$, $\mathrm{CCE}(\overline{Q},\underline{Q})$ returns a distribution $\pi \in \Delta_{\mathcal{A}\times\mathcal{B}}$ such that

$$\mathbb{E}_{(a,b)\sim\pi}\overline{Q}(a,b) \geq \sup_{a^*}\mathbb{E}_{(a,b)\sim\pi}\overline{Q}(a^*,b),$$

$$\mathbb{E}_{(a,b)\sim\pi}\underline{Q}(a,b) \leq \inf_{b^*}\mathbb{E}_{(a,b)\sim\pi}\overline{Q}(a,b^*).$$

The players choose their actions in a potentially correlated way so that no one can benefit from unilateral unconditional deviation. Since Nash equilibrium is also a CCE and a Nash equilibrium always exists, a CCE therefore always exist. Moreover, CCE can be efficiently implemented by linear programming in polynomial time. We remark that the policies generated by CCE are in general correlated, and

executing such policies requires the cooperation of the two players (line 6).

Algorithm description. For clarity, we provide a schematic algorithm here (Algorithm 1) and defer the detail to the appendix (Algorithm 3). Besides the standard optimistic and pessimistic value estimates update $\overline{Q}_h(s,a,b)$, $\overline{V}_h(s)$, $\underline{Q}_h(s,a,b)$, $\underline{V}_h(s)$, and the reference value functions $\overline{V}_h^{\mathrm{ref}}(s)$, $\underline{V}_h^{\mathrm{ref}}(s)$, the algorithm keeps multiple different accumulators to facilitate the update: 1) $N_h(s,a,b)$ and $\check{N}_h(s,a,b)$ are used to keep the total visit number and the visits counting for the current stage with respect to (s,a,b,h), respectively. 2) Intra-stage accumulators are used in the latest stage and are reset at the beginning of each stage. 3) The global accumulators are used for the samples in all stages: All accumulators are initialized to 0 at the beginning of the algorithm. The details of the accumulators are deferred to Appendix A.

The algorithm set $\iota = \log(2/\delta)$, $\beta = O(1/H)$ and $N_0 = c_4 SABH^5\iota/\beta^2$ for some sufficiently large universal constant c_4, denoting the number of visits required to learn β-accurate pair of reference value functions.

Certified policy. Based on the policy trajectories collected from Algorithm 3, we construct an output policy profile $(\mu^{\mathrm{out}}, \nu^{\mathrm{out}})$ that we will show is an approximate NE. For any step $h \in [H]$, an episode $k \in [K]$ and any state, we let $\mu_h^k(\cdot|s) \in \Delta(\mathcal{A})$ and $\nu_h^k(\cdot|s) \in \Delta(\mathcal{B})$ be the distribution prescribed by Algorithm 3 at this step. Let $\check{N}_h^k(s)$ be the value $\check{N}_h^k(s)$ at the beginning of the k-th episode. Our construction of the output policy μ^{out} is presented in Algorithm 2 (whereas the certified policy ν^{out} of the min-player can be obtained similarly), which follows the "certified policies" introduced in [Bai & Jin, 2020a]. We remark that the episode index from the previous stage is uniformly sampled in our algorithm while the certified policies in [Bai & Jin, 2020a] uses a weighted mixture.

Algorithm 2 Certified policy μ^{out} (max-player version)

1: Sample $k \leftarrow \mathrm{Unif}([K])$.
2: **for** step $h \leftarrow 1, \ldots, H$ **do**
3: Receive s_h, and take action $a_h \sim \mu_h^k(\cdot|s_h)$.
4: Observe b_h.
5: Sample $j \leftarrow \mathrm{Unif}([N_h^k(s_h, a_h, b_h)])$.
6: Set $k \leftarrow \check{\ell}_{h,j}^k$.
7: **end for**

4. Theoretical Analysis

4.1. Main Result

In this subsection, we present the main theoretical result for Algorithm 3. The following theorem presents the sample complexity guarantee for Algorithm 3 to learn a near-optimal Nash equilibrium policy in two-player zero-sum Markov games, which improves the best-known model-free algorithms in the same setting.

Theorem 4.1. *For any $\delta \in (0,1)$, let the agents run Algorithm 3 for K episodes with $K \geq \widetilde{O}(H^3SAB/\epsilon^2)$. Then, with probability at least $1-\delta$, the output policy $(\mu^{\mathrm{out}}, \nu^{\mathrm{out}})$ of Algorithm 2 is an ϵ-approximate Nash equilibrium.*

Compared to the lower bound $\Omega(H^3S(A+B)/\epsilon^2)$ on the sample complexity to find a near-optimal Nash equilibrium [Bai & Jin, 2020b], the sample complexity in Theorem 4.1 is minimax-optimal on the dependence of H, S and ϵ. This is the first result that establishes such optimality for model-free algorithms, although model-based algorithms have been shown to achieve such optimality in the past [Liu et al., 2021].

We also note that the result in Theorem 4.1 is not tight on the dependence on the cardinality of actions A, B. Such a gap has been closed by popular V-learning algorithms [Liu et al., 2021; Mao et al., 2022], which achieve the sample complexity of $O(H^5S(A+B)/\epsilon^2)$ [Mao et al., 2022]. Clearly, V-learning achieves a tight dependence on A, B, but suffers from worse horizon dependence on H. More specifically, one H factor is due to the nature of implementing the adversarial bandit subroutine in exchange for a better action dependence $A+B$. The other H factor could potentially be improved via the reference-advantage decomposition technique that we adopt here for our Q-learning algorithm. We leave this promising yet challenging direction as a future study.

4.2. Proof Outline

In this section, we present the proof sketch of Theorem 4.1, and defer all the details to the appendix.

Our main technical development lies in establishing a few new properties on the cumulative occurrence of the large V-gap and the cumulative bonus term, which enable the upper-bounding of several new error terms arising due to the incorporation of the new min-gap based reference-advantage decomposition technique. These properties have not been established for the single-agent RL with such a technique, because our properties are established for policies generated by the CCE oracle in zero-sum Markov games. Further, we perform a more refined analysis for both the optimistic and pessimistic accumulative bonus terms in order to obtain the desired result.

For certain functions, we use the superscript k to denote the value of the function at the beginning of the k-th episode, and use the superscript $K+1$ to denote the value of the function after all K episodes are played. For instance, we denote $N_h^k(s,a,b)$ as the value of $N_h(s,a,b)$ at the beginning of the k-th episode, and $N_h^{K+1}(s,a,b)$ to denote the

total number of visits of (s,a,b) at step h after K episodes. When h and k are clear from the context, we omit the subscript h and superscript k for notational convenience. For example, we use ℓ_i and $\check{\ell}_i$ to denote $\ell_{h,i}^k$ and $\check{\ell}_{h,i}^k$ when h and k are obvious.

In the next four steps, we strive to bound the difference between optimistic and pessimistic value functions $\frac{1}{K}\sum_{k=1}^K(\overline{V}_1^k - \underline{V}_1^k)(s_1)$, which is shown to upper bound our final goal $V_1^{\dagger,\nu^{\text{out}}}(s_1) - V_1^{\mu^{\text{out}},\dagger}(s_1)$ in the final step (Lemma 4.6).

Step I: We show that the Nash equilibrium (action-)value functions are always bounded between the optimistic and pessimistic (action-)value functions.

Lemma 4.2. *With high probability, it holds that for any s,a,b,k,h,*

$$\underline{Q}_h^k(s,a,b) \le Q_h^*(s,a,b) \le \overline{Q}_h^k(s,a,b),$$
$$\underline{V}_h^k(s) \le V_h^*(s) \le \overline{V}_h^k(s).$$

Our new technical development lies in proving the inequality with respect to the action-value function, whose update rule features the min-gap reference-advantage decomposition in two-player zero-sum Markov game.

The proof is by induction. We will focus on the optimistic (action-)value function and the other direction for pessimistic (action-)value function can be proved similarly. Suppose the two inequalities hold in episode k. We first establish the inequality for action-value function, and then prove the inequality for value functions. Based on the update rule of the optimistic action-value functions (line 8-9 in Algorithm 1, and line 12 in Algorithm 3), the action-value function is determined by the first two non-trial terms and last trivial term. While the first term is shown to upper bound the action-value function at Nash equilibrium $Q_h^*(s,a,b)$, we make the effort to showcase that the second term involving the min-gap based reference-advantage decomposition also upper bounds $Q_h^*(s,a,b)$. Since the optimistic action-value function takes the minimum of the three terms, we conclude that the optimistic action-value function in episode $k+1$ also satisfy the inequality. The proof of the inequality for value function (second inequality in Lemma 4.2) is based on the property of the policy distribution output by the CCE oracle.

Note that the optimistic (or pessimistic) action-value function is non-increasing (or non-decreasing) with respect to the iteration number k. However, the optimistic and the pessimistic value functions do not necessarily preserve such monotonic property due to the nature of the CCE oracle. This motivates our design of the min-gap based reference-advantage decomposition.

Step II: We show that the reference value function can be learned with bounded sample complexity in the following lemma.

Lemma 4.3. *With high probability, it holds that*

$$\sum_{k=1}^K \mathbf{1}\{\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \ge \epsilon\} \le O(SABH^5\iota/\epsilon^2)$$

We show that in the two-player zero-sum Markov game, the occurrence of the large V-gap, induced by the policy generated by the CCE oracle, is bounded independent of the number of episodes K. Our new development in proving this lemma lies in handling an additional martingale difference arising due to the CCE oracle.

In order to extract the best pair of optimistic and pessimistic value functions, a key novel min-gap based reference-advantage decomposition is proposed (see Section 3), based on which we pick up the pair of optimistic and pessimistic value functions whose gap is the smallest in the history (line 16-20 in Algorithm 1 and line 17-20 in Algorithm 3). The motivation is based on the observation mentioned in step I, and the latest pair of optimistic and pessimistic value functions does not necessarily have the minimum gap in this history. By the selection of the reference value functions, Lemma 4.3 with ϵ set to β, and the definition of N_0 (see Section 3 Algorithm description), we have the following corollary.

Corollary 4.4. *Conditioned on the successful events of Proposition 4.2 and Lemma 4.3, for every state s, we have*

$$n_h^k(s) \ge N_0 \Longrightarrow \overline{V}_h^{\text{ref},k}(s) - \underline{V}_h^{\text{ref},k}(s) \le \beta.$$

Step III: We bound $\sum_{k=1}^K(\overline{V}_1^k - \underline{V}_1^k)(s_1)$. Compared to single-agent RL, the CCE oracle leads to a possibly mixed policy and we need to bound the additional term due to the CCE oracle.

For ease of exposition, define $\Delta_h^k = (\overline{V}_h^k - \underline{V}_h^k)(s_h^k)$, and martingale difference $\zeta_h^k = \Delta_h^k - (\overline{Q}_h^k - \underline{Q}_h^k)(s_h^k,a_h^k,b_h^k)$. Note that $n_h^k = N_h^k(s_h^k,a_h^k,b_h^k)$ and $\check{n}_h^k = \check{N}_h^k(s_h^k,a_h^k,b_h^k)$ when $N_h^k(s_h^k,a_h^k,b_h^k) \in \mathcal{L}$. Following the update rule, we have (omitting the detail)

$$\begin{aligned}\Delta_h^k &= \zeta_h^k + (\overline{Q}_h^k - \underline{Q}_h^k)(s_h^k,a_h^k,b_h^k)\\ &\le \zeta_h^k + H\mathbf{1}\{n_h^k = 0\} + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\Delta_{h+1}^{\check{\ell}_i} + \Lambda_{h+1}^k,\end{aligned}$$

where the definition of Λ_{h+1}^k is provided in the appendix.

Summing over $k \in [K]$, we have

$$\sum_{k=1}^K \Delta_h^k$$

$$\leq \sum_{k=1}^{K} \zeta_h^k + \sum_{k=1}^{K} H\mathbf{1}\{n_h^k = 0\} + \sum_{k=1}^{K} \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \Delta_{h+1}^{\check{\ell}_{h,i}^k} + \sum_{k=1}^{K} \Lambda_{h+1}^k$$
$$\leq \sum_{k=1}^{K} \zeta_h^k + SABH^2 + (1 + \frac{1}{H}) \sum_{k=1}^{K} \Delta_{h+1}^k + \sum_{k=1}^{K} \Lambda_{h+1}^k,$$

where in the last inequality, we use the pigeon-hole argument for the second term, and the third term is due to the $(1 + 1/H)$ growth rate of the length of the stages.

Before we proceed, we briefly discuss several differences between our analysis for zero-sum game and single-agent RL. First, we care about the value difference between optimistic and pessimistic value functions in two-player zero-sum game instead of the value different between optimistic value function and the value function when executing policy π_k in single-agent RL. Second, additional martingale difference $\{\zeta_h^k\}_{(h,k)\in[K]\times[H]}$ shows up in two-player zero-sum game due to the fact that the CCE oracle in general output a mixed policy.

Iterating over $h = H, H-1, \ldots, 1$ gives

$$\sum_{k=1}^{K} \Delta_1^k \leq \mathcal{O}\left(SABH^3 + \sum_{h=1}^{H}\sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} \zeta_h^k + \sum_{h=1}^{H}\sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} \Lambda_{h+1}^k \right).$$

As pointed out earlier, the additional term $\sum_{h=1}^{H}\sum_{k=1}^{K}(1 + \frac{1}{H})^{h-1}\zeta_h^k$ is new in the two-player zero-sum Markov game, which can be bounded by Azuma-Hoeffding's inequality. I.e., it holds that with probability at least $1 - T\delta$,

$$\sum_{h=1}^{H}\sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} \zeta_h^k \leq \mathcal{O}(\sqrt{H^2 T\iota}),$$

which turns out to be a lower-order term compared to $\sum_{h=1}^{H}\sum_{k=1}^{K}(1 + \frac{1}{H})^{h-1}\Lambda_{h+1}^k$.

Step IV: We bound $\sum_{h=1}^{H}\sum_{k=1}^{K}(1 + \frac{1}{H})^{h-1}\Lambda_{h+1}^k$ in the following lemma.

Lemma 4.5. *With high probability, it holds that*

$$\sum_{h=1}^{H}\sum_{k=1}^{K} (1 + \frac{1}{H})^{h-1} \Lambda_{h+1}^k = O\left(\sqrt{SABH^2\iota} + H\sqrt{T\iota}\log T + S^2(AB)^{\frac{3}{2}} H^8 \iota T^{\frac{1}{4}} \right).$$

We capture the accumulative error of the bonus terms $\sum_{h=1}^{H}\sum_{k=1}^{K}(1 + \frac{1}{H})^{h-1}(\overline{\beta}_{h+1}^k + \underline{\beta}_{h+1}^k)$ in the expression $\sum_{h=1}^{H}\sum_{k=1}^{K}(1 + \frac{1}{H})^{h-1}\Lambda_{h+1}^k$. Since we first implement the reference-advantage decomposition technique in the two-player zero-sum game, our accumulative bonus term is much more challenging to analyze than the existing Q-learning algorithms for games. Compared to the analysis for the model-free algorithm with reference-advantage decomposition in single-RL [Zhang et al., 2020b], our analysis features the following **new developments**. First, we need to bound both the optimistic and pessimistic accumulative bonus terms, and the analysis is not identical. Second, the analysis of the optimistic accumulative bonus term differs due to the CCE oracle and the new min-gap base reference-advantage decomposition for two-player zero-sum Markov game.

Final step. We build connection between the certified policy generated by Algorithm 2, and the difference between the optimistic and pessimistic value functions $\frac{1}{K}\sum_{k=1}^{K}(\overline{V}_1^k - \underline{V}_1^k)(s_1)$.

Lemma 4.6. *Let $(\mu^{\text{out}}, \nu^{\text{out}})$ be the output policy induced by the certified policy algorithm (Algorithm 2), then, we have*

$$V_1^{\dagger,\nu^{\text{out}}}(s_1) - V_1^{\mu^{\text{out}},\dagger}(s_1) \leq \frac{1}{K}\sum_{k=1}^{K}(\overline{V}_1^k - \underline{V}_1^k)(s_1).$$

Finally, combining all steps, we conclude that with high probability,

$$V_1^{\dagger,\nu^{\text{out}}}(s_1) - V_1^{\mu^{\text{out}},\dagger}(s_1) \leq \frac{1}{K}\sum_{k=1}^{K}\Delta_h^k = O\left(\frac{H^3 SAB}{\epsilon^2}\right).$$

5. Conclusion

In this paper, we proposed a new model-free algorithm Q-learning with min-gap based reference-advantage decomposition for two-player zero-sum Markov games, which improved the existing results and achieved a near-optimal sample complexity $O(H^3SAB/\epsilon^2)$ except for the AB term. Due to the nature of the CCE oracle employed in the algorithm, we designed a novel min-gap based reference-advantage decomposition to learn the pair of optimistic and pessimistic reference value functions whose value difference has the minimum gap in the history. An interesting future direction would be to study whether the horizon dependence could be further tightened in model-free V-learning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

The work of S. Feng and Y. Liang was supported in part by the U.S. National Science Foundation under the grants RINGS-2148253, CCF-1900145 and AI-EDGE Institute CNS-2112471.

References

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2008.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020a.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, 2020.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020.

Chen, F., Mei, S., and Bai, Y. Unified Algorithms for RL with Decision-Estimation Coefficients: No-Regret, PAC, and Reward-Free Learning. *arXiv e-prints*, 2022.

Chen, Z., Zhou, D., and Gu, Q. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, 2022.

Cui, Q., Zhang, K., and Du, S. S. Breaking the Curse of Multiagents in a Large State Space: RL in Markov Games with Independent Linear Function Approximation. *arXiv e-prints*, 2023.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Daskalakis, C., Golowich, N., and Zhang, K. The Complexity of Markov Equilibrium in Stochastic Games. *arXiv e-prints*, 2022.

Filar, J. and Vrieze, K. *Competitive Markov Decision Processes*. Springer, 1997.

Hansen, T. D., Miltersen, P. B., and Zwick, U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J. ACM*, 60(1), feb 2013. ISSN 0004-5411.

Hu, J. and Wellman, M. P. Nash Q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4(null): 1039–1069, dec 2003. ISSN 1532-4435.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. Towards general function approximation in zero-sum markov games. In *International Conference on Learning Representations*, 2022.

Jia, Z., Yang, L. F., and Wang, M. Feature-Based Q-Learning for Two-Player Stochastic Games. *arXiv e-prints*, 2019.

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

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning – a simple, efficient, decentralized algorithm for multiagent RL. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022a. URL https://openreview.net/forum?id=Bx-evj5k6x9.

Jin, C., Liu, Q., and Yu, T. The power of exploiter: Provable multi-agent RL in large state spaces. In *Proceedings of the 39th International Conference on Machine Learning*, 2022b.

Li, G., Chi, Y., Wei, Y., and Chen, Y. Minimax-optimal multi-agent RL in markov games with a generative model. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=W8nyVJruVg.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, 1994.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Mao, W. and Başar, T. Provably Efficient Reinforcement Learning in Decentralized General-Sum Markov Games. *arXiv e-prints*, 2021.

Mao, W., Yang, L., Zhang, K., and Basar, T. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Ni, C., Song, Y., Zhang, X., Ding, Z., Jin, C., and Wang, M. Representation learning for low-rank general-sum markov games. In *The Eleventh International Conference on Learning Representations*, 2023.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. *arXiv e-prints*, 2016.

Shapley, L. S. Stochastic games. *Proceedings of the National Academy of Sciences*, 39:1095 – 1100, 1953.

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

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

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T. P., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017.

Song, Z., Mei, S., and Bai, Y. When can we learn general-sum markov games with a large number of players sample-efficiently? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=6MmiS0HUJHR.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T. P., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. Grandmaster level in starcraft II using multi-agent reinforcement learning. *Nature*, pp. 1–5, 2019.

Wang, Y., Liu, Q., Bai, Y., and Jin, C. Breaking the Curse of Multiagency: Provably Efficient Decentralized Multi-Agent RL with Function Approximation. *arXiv e-prints*, 2023.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Linear Last-iterate Convergence in Constrained Saddle-point Optimization. *arXiv e-prints*, 2020.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

Xiong, W., Zhong, H., Shi, C., Shen, C., and Zhang, T. A self-play posterior sampling algorithm for zero-sum Markov games. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Zhan, W., Lee, J. D., and Yang, Z. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhang, K., Kakade, S. M., Başar, T., and Yang, L. F. Model-based multi-agent RL in zero-sum markov games with near-optimal sample complexity. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020a.

Zhang, Z., Zhou, Y., and Ji, X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020b.

Supplementary Materials

A. Details of Algorithm 1

Algorithm 3 Q-learning with min-gap based reference-advantage decomposition

Initialize: Set all accumulators to 0. For all $(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$, set $\overline{V}_h(s), \overline{Q}_h(s,a,b)$ to $H-h+1$, set $\overline{V}^{\mathrm{ref}}_h(s)$ to H, set $\underline{V}_h(s), \underline{Q}_h(s,a,b), \underline{V}^{\mathrm{ref}}_h(s,a,b)$ to 0; and
let $\pi_h(s) \sim \mathrm{Unif}(\mathcal{A})\times\mathrm{Unif}(\mathcal{B})$, $\Delta(s,h)=H$, $\widetilde{\overline{V}}_h(s_h)=H$, $\widetilde{\underline{V}}_h(s_h)=0$.
for episodes $k \leftarrow 1,2,\ldots,K$ **do**
  Observe s_1.
  for $h \leftarrow 1,2,\ldots,H$ **do**
    Take action $(a_h,b_h) \leftarrow \pi_h(s_h)$, receive $r_h(s_h,a_h,b_h)$, and observe s_{h+1}.
    Update accumulators $n := N_h(s_h,a_h,b_h) \overset{+}{\leftarrow} 1$, $\check{n} := \check{N}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} 1$ and (5)-(9).
    if $n \in \mathcal{L}$ **then**
      $\gamma \leftarrow 2\sqrt{\frac{H^2}{\check{n}}\iota}$.
      $\overline{\beta} \leftarrow c_1\sqrt{\frac{\overline{\sigma}^{\mathrm{ref}}/n-(\overline{\mu}^{\mathrm{ref}}/n)^2}{n}\iota} + c_2\sqrt{\frac{\check{\overline{\sigma}}/\check{n}-(\check{\overline{\mu}}/\check{n})^2}{\check{n}}\iota} + c_3(\frac{H\iota}{n}+\frac{H\iota}{\check{n}}+\frac{H\iota^{3/4}}{n^{3/4}}+\frac{H\iota^{3/4}}{\check{n}^{3/4}})$.
      $\underline{\beta} \leftarrow c_1\sqrt{\frac{\underline{\sigma}^{\mathrm{ref}}/n-(\underline{\mu}^{\mathrm{ref}}/n)^2}{n}\iota} + c_2\sqrt{\frac{\check{\underline{\sigma}}/\check{n}-(\check{\underline{\mu}}/\check{n})^2}{\check{n}}\iota} + c_3(\frac{H\iota}{n}+\frac{H\iota}{\check{n}}+\frac{H\iota^{3/4}}{n^{3/4}}+\frac{H\iota^{3/4}}{\check{n}^{3/4}})$.
      $\overline{Q}_h(s_h,a_h,b_h) \leftarrow \min\{r_h(s_h,a_h,b_h)+\frac{\check{\overline{v}}}{\check{n}}+\gamma, r_h(s_h,a_h,b_h)+\frac{\overline{\mu}^{\mathrm{ref}}}{n}+\frac{\check{\overline{\mu}}}{\check{n}}+\overline{\beta}, \overline{Q}_h(s_h,a_h,b_h)\}$.
      $\underline{Q}_h(s_h,a_h,b_h) \leftarrow \max\{r_h(s_h,a_h,b_h)+\frac{\check{\underline{v}}}{\check{n}}-\gamma, r_h(s_h,a_h,b_h))+\frac{\underline{\mu}^{\mathrm{ref}}}{n}+\frac{\check{\underline{\mu}}}{\check{n}}-\underline{\beta}, \underline{Q}_h(s_h,a_h,b_h)\}$.
      $\pi_h(s_h) \leftarrow \mathrm{CCE}(\overline{Q}(s_h,\cdot,\cdot), \underline{Q}_h(s_h,\cdot,\cdot))$.
      $\overline{V}_h(s_h) \leftarrow \mathbb{E}_{(a,b)\sim\pi_h(s_h)}\overline{Q}_h(s_h,a,b)$, and $\underline{V}_h(s_h) \leftarrow \mathbb{E}_{(a,b)\sim\pi_h(s_h)}\underline{Q}_h(s_h,a,b)$.
      Reset all intra-stage accumulators to 0.
      if $\overline{V}_h(s_h)-\underline{V}_h(s_h) < \Delta(s,h)$ **then**
        $\Delta(s,h) = \overline{V}_h(s_h)-\underline{V}_h(s_h)$.
        $\widetilde{\overline{V}}_h(s_h) = \overline{V}_h(s_h)$, $\widetilde{\underline{V}}_h(s_h) = \underline{V}_h(s_h)$.
      end if
    end if
    if $\sum_{a,b} N_h(s_h,a,b) = N_0$ **then**
      $\overline{V}^{\mathrm{ref}}_h(s_h) \leftarrow \widetilde{\overline{V}}_h(s_h)$, $\underline{V}^{\mathrm{ref}}_h(s_h) \leftarrow \widetilde{\underline{V}}_h(s_h)$.
    end if
  end for
end for

Algorithm description. Let c_1, c_2, c_3 be some sufficiently large universal constants so that the concentration inequalities can be applied in the analysis. Besides the standard optimistic and pessimistic value estimates $\overline{Q}_h(s,a,b)$, $\overline{V}_h(s)$, $\underline{Q}_h(s,a,b)$, $\underline{V}_h(s)$, and the reference value functions $\overline{V}^{\mathrm{ref}}_h(s)$, $\underline{V}^{\mathrm{ref}}_h(s)$, the algorithm keeps multiple different accumulators to facilitate the update: 1) $N_h(s,a,b)$ and $\check{N}_h(s,a,b)$ are used to keep the total visit number and the visits counting for the current stage with respect to (s,a,b,h), respectively. 2) Intra-stage accumulators are used in the latest stage and are reset at the beginning of each stage. The update rule of the intra-stage accumulators are as follows:

$$\check{\overline{v}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} \overline{V}_{h+1}(s_{h+1}), \quad \check{\underline{v}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} \underline{V}_{h+1}(s_{h+1}), \tag{5}$$

$$\check{\overline{\mu}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} \overline{V}_{h+1}(s_{h+1}) - \overline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}), \quad \check{\underline{\mu}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} \underline{V}_{h+1}(s_{h+1}) - \underline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}), \tag{6}$$

$$\check{\overline{\sigma}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} (\overline{V}_{h+1}(s_{h+1}) - \overline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}))^2, \check{\underline{\sigma}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} (\underline{V}_{h+1}(s_{h+1}) - \underline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}))^2. \tag{7}$$

3) The following global accumulators are used for the samples in all stages:

$$\overline{\mu}^{\mathrm{ref}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} \overline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}), \quad \underline{\mu}^{\mathrm{ref}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} \underline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}), \tag{8}$$

$$\overline{\sigma}^{\mathrm{ref}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} (\overline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}))^2, \quad \underline{\sigma}^{\mathrm{ref}}_h(s_h,a_h,b_h) \overset{+}{\leftarrow} (\underline{V}^{\mathrm{ref}}_{h+1}(s_{h+1}))^2. \tag{9}$$

All accumulators are initialized to 0 at the beginning of the algorithm. The algorithm set $\iota = \log(2/\delta)$, $\beta = O(1/H)$ and $N_0 = c_4 SABH^5/\beta^2$ for some sufficiently large universal constant c_4.

B. Comparison to Existing Algorithms

Compare to Optimistic Nash Q-learning [Bai et al., 2020]. The Optimistic Nash Q-learning is a model-free Q-learning algorithm for two-player zero-sum Markov games. The algorithm design differences between our algorithm and the optimistic Nash Q-learning is two-fold. First, we adopt the stage-based design instead of traditional Q-learning update $Q_{new} \leftarrow (1-\alpha)Q_{old} + \alpha(r+V)$. The optimistic Nash Q-learning updates the value function with a learning rate, while our algorithm adopts greedy update. We remark that both frameworks are viable, and in our opinion, the stage-based design is easier to follow and analyse. Second, we propose a novel min-gap based reference-advantage decomposition, a variance reduction technique, to further improve the sample complexity. Specifically, we use both the standard update rule and the advantage-based update rule in our action-value function (Q function) while the optimistic Nash Q-learning only uses the standard update rule.

Aside from the obvious distinction of the proofs caused by stage-based design, the main difference is the analysis for the advantage-based update rule, which does not show up in the optimistic Nash Q-learning. Due to the incorporation of the new min-gap based reference-advantage decomposition technique, several new error terms arise in our analysis. Our main development lies in establishing a few new properties on the cumulative occurrence of the large V-gap and the cumulative bonus term, which enable the upper-bounding of those new error terms. More specifically, as we explain in our proof outline in Section 4.2, our analysis include the following novel developments. (i) Step I shows that the Nash equilibrium (action-)value functions are always bounded between the optimistic and pessimistic (action-)value functions (see Lemma 4.3). Our new technical development here lies in proving the inequality with respect to the action-value function, whose update rule features the min-gap reference-advantage decomposition. (ii) Step II shows that the reference value can be learned with bounded sample complexity (see Lemma 4.4). Our new development here lies in handling an additional martingale difference arising due to the CCE oracle. (iii) In step IV, there are a few new developments. First, we need to bound both the optimistic and pessimistic accumulative bonus terms, and the analysis is more refined compared to that for single-agent RL. Second, the analysis of the optimistic accumulative bonus term need to handle the CCE oracle together with the new min-gap base reference-advantage decomposition for two-player zero-sum Markov game.

Compare to UCB-advantage [Zhang et al., 2020b]. The UCB-advantage is a model-free algorithm with reference-advantage decomposition for single-agent RL. Our main novel design idea lies in the **min-gap** based advantage reference value decomposition. Unlike the single-agent scenario, the optimistic (or pessimistic) value function in Markov games does not necessarily preserve the monotone property due to the nature of the CCE oracle. In order to obtain the "best" optimistic and pessimistic value function pair, we propose the key **min-gap** design to update the reference value functions as the pair of optimistic and pessimistic value functions whose value difference is the smallest (i.e., with the minimal gap) in the history. It turns out that such a design is critical to guarantee the provable sample efficiency.

For the proof techniques, there are the fundamental differences between single-agent RL and two-player zero-sum games. Thanks to the key min-gap based reference-advantage decomposition, we provide a new guarantee for the learned pair of reference value (Corollary 4.5) in the context of two-player zero-sum Markov games, which is crucial in obtaining an optimal horizon dependence.

C. Notations

For any function $f : \mathcal{S} \mapsto \mathbb{R}$, we use $P_{s,a,b}f$ and $(P_h f)(s,a,b)$ interchangeably. Define $\mathbb{V}(x,y) = x^\top(y^2) - (x^\top y)^2$ for two vectors of the same dimension, where y^2 is obtained by squaring each entry of y.

For ease of exposition, we define $\overline{\nu}_h^{\mathrm{ref},k} = \frac{\overline{\sigma}_h^{\mathrm{ref},k}}{n_h^k} - (\frac{\overline{\mu}_h^{\mathrm{ref},k}}{n_h^k})^2$, $\underline{\nu}_h^{\mathrm{ref},k} = \frac{\overline{\sigma}_h^{\mathrm{ref},k}}{n_h^k} - (\frac{\underline{\mu}_h^{\mathrm{ref},k}}{n_h^k})^2$ and $\check{\overline{\nu}}_h^k = \frac{\check{\overline{\sigma}}_h^k}{\check{n}_h^k} - (\frac{\check{\overline{\mu}}_h^k}{\check{n}_h^k})^2$, $\check{\underline{\nu}}_h^k = \frac{\check{\underline{\sigma}}_h^k}{\check{n}_h^k} - (\frac{\check{\underline{\mu}}_h^k}{\check{n}_h^k})^2$. Moreover, we define $\Delta_h^k = \overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)$ and $\zeta_h^k = \Delta_h^k - (\overline{Q}_h^k - \underline{Q}_h^k)(s_h^k, a_h^k, b_h^k)$. For convenience, we also define $\lambda_h^k(s) = \mathbf{1}\left\{n_h^k(s) < N_0\right\}$.

For certain functions, we use the superscript k to denote the value of the function at the beginning of the k-th episode, and use the superscript $K+1$ to denote the value of the function after all K episodes are played. For instance, we denote $N_h^k(s,a,b)$ as the value of $N_h(s,a,b)$ at the beginning of the k-th episode, and $N_h^{K+1}(s,a,b)$ to denote the total number of

visits of (s,a,b) at step h after K episodes. When it is clear from the context, we omit the subscript h and the superscript k for notational convenience. For example, we use ℓ_i and $\check{\ell}_i$ to denote $\ell_{h,i}^k$ and $\check{\ell}_{h,i}^k$ when it is obvious what values that the indices h and k take.

D. Proof of Theorem 4.1

In this section, we provide the proof of Theorem 4.1, which consists of four main steps and one final step. In order to provide a clear proof flow here, we defer the proofs of the main lemmas in these steps to later sections (i.e., Appendix E-Appendix H).

We start by replacing δ by $\delta/\mathrm{poly}(H,T)$, and it suffices to show the desired bound for $V_1^{\dagger,\nu^{\mathrm{out}}}(s_1) - V_1^{\mu^{\mathrm{out}},\dagger}(s_1)$ with probability $1-\mathrm{poly}(H,T)\delta$.

Step I: We show that the Nash equilibrium (action-)value functions are always bounded between the optimistic and pessimistic (action-)value functions.

Lemma D.1 (Restatement of Lemma 4.2)**.** *Let $\delta \in (0,1)$. With probability at least $1-2T(2H^2T^3+7)\delta$, it holds that for any s,a,b,k,h,*

$$\underline{Q}_h^k(s,a,b) \le Q_h^*(s,a,b) \le \overline{Q}_h^k(s,a,b),$$
$$\underline{V}_h^k(s) \le V_h^*(s) \le \overline{V}_h^k(s).$$

The proof of Lemma D.1 is provided in Appendix E. The **new technical development** lies in proving the inequality with respect to the action-value function, whose update rule features the min-gap reference-advantage decomposition.

Step II: We show that the occurrence of the large V-gap has bounded sample complexity independent of the number of episodes K.

Lemma D.2 (Restatement of Lemma 4.3)**.** *With probability $1-O(T\delta)$, it holds that*

$$\sum_{k=1}^{K} \mathbf{1}\{\overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k) \ge \epsilon\} \le O(SABH^5\iota/\epsilon^2).$$

The proof is provided in Appendix F.

By the selection of the reference value functions, Lemma D.2 with ϵ setting to β, and the definition of N_0, we have the following corollary.

Corollary D.3 (Restatement of Corollary 4.4)**.** *Conditioned on the successful events of Lemma D.1 and Lemma D.2, for every state s, we have*

$$n_h^k(s) \ge N_0 \Longrightarrow \overline{V}_h^{\mathrm{ref},k}(s) - \underline{V}_h^{\mathrm{ref},k}(s) \le \beta.$$

Step III: We bound $\sum_{k=1}^K(\overline{V}_1^k - \underline{V}_1^k)(s_1)$. Compared to single-agent RL, the CCE oracle leads to a possibly mixed policy and we need to bound the additional term due to the CCE oracle.

Recall the definition of $\Delta_h^k = \overline{V}_h^k(s_h^k) - \underline{V}_h^k(s_h^k)$ and $\zeta_h^k = \Delta_h^k - (\overline{Q}_h^k - \underline{Q}_h^k)(s_h^k,a_h^k,b_h^k)$. Following the update rule, we have

$$\begin{aligned}
\Delta_h^k &= \zeta_h^k + (\overline{Q}_h^k - \underline{Q}_h^k)(s_h^k,a_h^k,b_h^k)\\
&\le \zeta_h^k + H\mathbf{1}\{n_h^k=0\} + \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) - \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\\
&\quad + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(\overline{V}_{h+1}^{\check{\ell}_i} - \overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i})(s_{h+1}^{\check{\ell}_i}) - \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}(\underline{V}_{h+1}^{\check{\ell}_i} - \underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i})(s_{h+1}^{\check{\ell}_i}) + \overline{\beta}_h^k + \underline{\beta}_h^k\\
&\le \zeta_h^k + H\mathbf{1}\{n_h^k=0\} + \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i} - \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}
\end{aligned}$$

$$+\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} P_{s_h^k,a_h^k,b_h^k,h}(\overline{V}_{h+1}^{\check{\ell}_i}-\overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i})-\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} P_{s_h^k,a_h^k,b_h^k,h}(\underline{V}_{h+1}^{\check{\ell}_i}-\underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i})+2\overline{\beta}_h^k+2\underline{\beta}_h^k \tag{10}$$

$$\begin{aligned}
&=\zeta_h^k+H\mathbf{1}\{n_h^k=0\}+P_{s_h^k,a_h^k,b_h^k,h}\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}-\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)\\
&\quad-P_{s_h^k,a_h^k,b_h^k,h}\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}-\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)+P_{s_h^k,a_h^k,b_h^k,h}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\check{\ell}_i}-\underline{V}_{h+1}^{\check{\ell}_i}\right)\\
&\quad+2\overline{\beta}_h^k+2\underline{\beta}_h^k\\
&\le\zeta_h^k+H\mathbf{1}\{n_h^k=0\}+P_{s_h^k,a_h^k,b_h^k,h}\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}-\overline{V}_{h+1}^{\mathrm{REF}}\right)\\
&\quad-P_{s_h^k,a_h^k,b_h^k,h}\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}-\underline{V}_{h+1}^{\mathrm{REF}}\right)+P_{s_h^k,a_h^k,b_h^k,h}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\check{\ell}_i}-\underline{V}_{h+1}^{\check{\ell}_i}\right)+2\overline{\beta}_h^k+2\underline{\beta}_h^k \qquad (11)\\
&=\zeta_h^k+H\mathbf{1}\{n_h^k=0\}+\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\Delta_{h+1}^{\check{\ell}_i}+\Lambda_{h+1}^k, \qquad (12)
\end{aligned}$$

where we define

$$\begin{aligned}
\Lambda_{h+1}^k&=\psi_{h+1}^k+\xi_{h+1}^k+2\overline{\beta}_h^k+2\underline{\beta}_h^k,\\
\psi_{h+1}^k&=P_{s_h^k,a_h^k,b_h^k,h}\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(\overline{V}_{h+1}^{\mathrm{ref},\ell_i}-\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)-\left(\overline{V}_{h+1}^{\mathrm{REF}}-\underline{V}_{h+1}^{\mathrm{REF}}\right)\right),\\
\xi_{h+1}^k&=\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^{\check{\ell}_i}}\right)\left(\overline{V}_{h+1}^{\check{\ell}_i}-\underline{V}_{h+1}^{\check{\ell}_i}\right).
\end{aligned}$$

Here, (10) follows from the successful event of martingale concentration (29) and (43) in Lemma D.1, (11) follows from the fact that $\overline{V}_{h+1}^{\mathrm{ref},u}(s)$ (or $\underline{V}_{h+1}^{\mathrm{ref},u}(s)$) is non-increasing (or non-decreasing) in u, because $\overline{V}_h^{\mathrm{ref}}(s)$ (or $\underline{V}_h^{\mathrm{ref}}(s)$) for a pair (s,h) can only be updated once and the updated value is obviously greater (or less) than the initial value, and (12) follows from the definition of Λ_{h+1}^k defined above.

Taking the summation over $k\in[K]$ gives

$$\sum_{k=1}^K\Delta_h^k\le\sum_{k=1}^K\zeta_h^k+\sum_{k=1}^K H\mathbf{1}\{n_h^k=0\}+\sum_{k=1}^K\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\Delta_{h+1}^{\check{\ell}_{h,i}^k}+\sum_{k=1}^K\Lambda_{h+1}^k. \tag{13}$$

Note that $n_h^k\ge H$ if $N_h^k(s_h^k,a_h^k,b_h^k)\ge H$. Therefore $\sum_{k=1}^K\mathbf{1}\{n_h^k=0\}\le SABH$, and

$$\sum_{k=1}^K H\mathbf{1}\{n_h^k=0\}\le SABH^2. \tag{14}$$

Now we focus on the term $\sum_{k=1}^K\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\Delta_{h+1}^{\check{\ell}_{h,i}^k}$. The following lemma is useful.

Lemma D.4. *For any $j\in[K]$, we have $\sum_{k=1}^K\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\mathbf{1}\{j=\check{\ell}_{h,i}^k\}\le 1+\frac{1}{H}$.*

Proof. Fix an episode j. Note that $\sum_{i=1}^{\check{n}_h^k}\mathbf{1}\{j=\check{\ell}_{h,i}^k\}=1$ if and only if $(s_h^j,a_h^j,b_h^j)=(s_h^k,a_h^k,b_h^k)$ and (j,h) falls in the previous stage that (k,h) falls in with respect to (s_h^k,a_h^k,b_h^k,h). Define $\mathcal{K}=\{k\in[K]:\sum_{i=1}^{\check{n}_h^k}\mathbf{1}\{j=\check{\ell}_{h,i}^k\}=1\}$. Then

every element $k \in \mathcal{K}$ has the same value of $\check{n}_h^k$, i.e., there exists an integer $N_j > 0$ such that $\check{n}_h^k = N_j$ for all $k \in \mathcal{K}$. By the definition of stages, $|\mathcal{K}| \le (1+\frac{1}{H})N_j$. Therefore, for any j, we have $\sum_{k=1}^K \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} \mathbf{1}\{j = \check{\ell}_{h,i}^k\} \le (1+\frac{1}{H})$. □

By Lemma D.4, we have

$$\begin{aligned}\sum_{k=1}^K \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} \Delta_{h+1}^{\check{\ell}_{h,i}^k} &= \sum_{k=1}^K \frac{1}{\check{n}_h^k}\sum_{j=1}^K \Delta_{h+1}^j \sum_{i=1}^{\check{n}_h^k} \mathbf{1}\{j = \check{\ell}_{h,i}^k\} \\ &= \sum_{j=1}^K \Delta_{h+1}^j \sum_{k=1}^K \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} \mathbf{1}\{j = \check{\ell}_{h,i}^k\} \\ &\le (1+\frac{1}{H})\sum_{k=1}^K \Delta_{h+1}^k. \end{aligned} \tag{15}$$

Combining (13), (14) and (15), we have

$$\sum_{k=1}^K \Delta_h^k \le SABH^2 + (1+\frac{1}{H})\sum_{k=1}^K \Delta_{h+1}^k + \sum_{k=1}^K \Lambda_{h+1}^k.$$

Iterating over $h = H, H-1, \ldots, 1$ gives

$$\sum_{k=1}^K \Delta_1^k \le \mathcal{O}\left(SABH^3 + \sum_{h=1}^H\sum_{k=1}^K (1+\frac{1}{H})^{h-1}\zeta_h^k + \sum_{h=1}^H\sum_{k=1}^K (1+\frac{1}{H})^{h-1}\Lambda_{h+1}^k\right).$$

By Azuma's inequality, it holds that with probability at least $1 - T\delta$,

$$\sum_{k=1}^K \Delta_1^k \le \mathcal{O}\left(SABH^3 + \sqrt{H^2T\iota} + \sum_{h=1}^H\sum_{k=1}^K (1+\frac{1}{H})^{h-1}\Lambda_{h+1}^k\right). \tag{16}$$

Step IV: We bound $\sum_{h=1}^H\sum_{k=1}^K (1+\frac{1}{H})^{h-1}\Lambda_{h+1}^k$ in the following lemma.

Lemma D.5 (Restatement of Lemma 4.5)**.** *With probability at least $1 - O(H^2T^4)\delta$, it holds that*

$$\sum_{h=1}^H\sum_{k=1}^K (1+\frac{1}{H})^{h-1}\Lambda_{h+1}^k = O\left(\sqrt{SABH^2T\iota} + H\sqrt{T\iota}\log T + S^2(AB)^{\frac{3}{2}}H^8\iota^{\frac{3}{2}}T^{\frac{1}{4}}\right).$$

The proof of Lemma D.5 is provided in Appendix G.

Final step: We show the value difference induced by the certified policies is bounded, as summarized in the next lemma.

Lemma D.6 (Restatement of Lemma 4.6)**.** *Conditioned on the successful event of Lemma D.1, let $(\mu^{\text{out}}, \nu^{\text{out}})$ be the output policy induced by the certified policy algorithm (Algorithm 2). Then we have*

$$V_1^{\dagger,\nu^{\text{out}}}(s_1) - V_1^{\mu^{\text{out}},\dagger}(s_1) \le \frac{1}{K}\sum_{k=1}^K (\overline{V}_1^k - \underline{V}_1^k)(s_1).$$

The proof of Lemma D.6 is provided in Appendix H.

Combining (16), Lemma D.5 and Lemma D.6, and taking the union bound over all probability events, we conclude that with probability at least $1 - O(H^2T^4)\delta$, it holds that

$$V_1^{\dagger,\nu^{\text{out}}}(s_1) - V_1^{\mu^{\text{out}},\dagger}(s_1) \le \frac{1}{K}O\left(\sqrt{SABH^2T\iota} + H\sqrt{T\iota}\log T + S^2(AB)^{\frac{3}{2}}H^8\iota^{\frac{3}{2}}T^{\frac{1}{4}}\right), \tag{17}$$

which gives the desired result.

E. Proof of Lemma D.1 (Step I)

The proof is by induction on k. We establish the inequalities for the optimistic action-value and value functions in **step i**, and the inequalities for the pessimistic counterparts in **step ii**.

Step i: We establish the inequality for the optimistic action-value and value functions in the following.

It is clear that the conclusion holds for the based case with $k = 1$. For $k \geq 2$, assume $Q_h^*(s,a,b) \leq \overline{Q}_h^u(s,a,b)$ and $V_h^*(s) \leq \overline{V}_h^u(s)$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $u \in [1,k]$. Fix tuple (s,a,b,h). We next show that the conclusion holds for $k+1$.

First, we show the inequality with respect to the action-value function. If $\overline{Q}_h(s,a,b), \overline{V}_h(s)$ are not updated in the k-th episode, then

$$\begin{aligned}
&Q_h^*(s,a,b) \leq \overline{Q}_h^k(s,a,b) = \overline{Q}_h^{k+1}(s,a,b),\\
&V_h^*(s) \leq \overline{V}_h^k(s) = \overline{V}_h^{k+1}(s).
\end{aligned}$$

Otherwise, we have

$$\overline{Q}_h^{k+1}(s,a,b) \leftarrow \min\left\{r_h(s,a,b) + \frac{\check{\overline{v}}}{\check{n}} + \gamma, r_h(s,a,b) + \frac{\overline{\mu}^{\mathrm{ref}}}{n} + \frac{\check{\overline{\mu}}}{\check{n}} + \overline{\beta}, \overline{Q}_h^k(s,a,b)\right\}.$$

Besides the last term, there are two non-trivial cases.

For the first case, by Hoeffding's inequality, with probability at least $1-\delta$ it holds that

$$\begin{aligned}
\overline{Q}_h^{k+1}(s,a,b) &= r_h(s,a,b) + \frac{\check{\overline{v}}}{\check{n}} + \gamma\\
&= r_h(s,a,b) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) + 2\sqrt{\frac{H^2}{\check{n}}\iota}\\
&\geq r_h(s,a,b) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} V_{h+1}^*(s_{h+1}^{\check{\ell}_i}) + 2\sqrt{\frac{H^2}{\check{n}}\iota} \qquad (18)\\
&\geq r_h(s,a,b) + (P_h V_{h+1}^*)(s,a,b) \qquad (19)\\
&= Q_h^*(s,a,b),
\end{aligned}$$

where (18) follows from the induction hypothesis $\overline{V}_{h+1}^u(s) \geq V^*(s)$ for all $u \in [k]$, and (19) follows from Azuma-Hoeffding's inequality.

For the second case, we have

$$\begin{aligned}
&\overline{Q}_h^{k+1}(s,a,b) = r_h(s,a,b) + \frac{\overline{\mu}^{\mathrm{ref}}}{n} + \frac{\check{\overline{\mu}}}{\check{n}} + \overline{\beta}\\
&= r_h(s,a,b) + \frac{1}{n}\sum_{i=1}^{n} \overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(\overline{V}_{h+1}^{\check{\ell}_i} - \overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)(s_{h+1}^{\check{\ell}_i}) + \overline{\beta}\\
&= r_h(s,a,b) + \left(P_h\left(\frac{1}{n}\sum_{i=1}^{n}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)\right)(s,a,b) + \left(P_h\left(\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(\overline{V}_{h+1}^{\check{\ell}_i} - \overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)\right)\right)(s,a,b)\\
&\qquad + \chi_1 + \chi_2 + \overline{\beta}\\
&\geq r_h(s,a,b) + \left(P_h\left(\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\overline{V}_{h+1}^{\check{\ell}_i}\right)\right)(s,a,b) + \chi_1 + \chi_2 + \overline{\beta} \qquad (20)\\
&\geq r_h(s,a,b) + \left(P_h V_{h+1}^*\right)(s,a,b) + \chi_1 + \chi_2 + \overline{\beta} \qquad (21)\\
&= \overline{Q}_h^*(s,a,b) + \chi_1 + \chi_2 + \overline{\beta},
\end{aligned}$$

where

$$\chi_1(k,h) = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{V}_h^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) - \left(P_h\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)(s,a,b)\right),$$
$$\overline{W}_{h+1}^{\ell} = \overline{V}_{h+1}^{\ell} - \overline{V}_{h+1}^{\mathrm{ref},\ell}$$
$$\chi_2(k,h) = \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(\overline{W}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \left(P_h\overline{W}_{h+1}^{\check{\ell}_i}\right)(s,a,b)\right).$$

Here, (20) follows from the fact that $\overline{V}_{h+1}^{\mathrm{ref},u}(s)$ is non-increasing in u (since $\overline{V}_h^{\mathrm{ref}}(s)$ for a pair (s,h) can only be updated once and the updated value is obviously smaller than the initial value H), and (21) follows from the the induction hypothesis $\overline{V}_{h+1}^{k}(s) \geq V_{h+1}^{*}(s)$.

By Lemma I.2 with $\epsilon = \frac{1}{T^2}$, with probability at least $1 - 2(H^2T^3+1)\delta$ it holds

$$|\chi_1(k,h)| \leq 2\sqrt{\frac{\sum_{i=1}^{n}\mathbb{V}(P_{s,a,b,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i})\iota}{n^2}} + \frac{2\sqrt{\iota}}{Tn} + \frac{2H\iota}{n}, \tag{22}$$

$$|\chi_2(k,h)| \leq 2\sqrt{\frac{\sum_{i=1}^{\check{n}}\mathbb{V}(P_{s,a,b,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i})\iota}{\check{n}^2}} + \frac{2\sqrt{\iota}}{T\check{n}} + \frac{2H\iota}{\check{n}}. \tag{23}$$

Lemma E.1. *With probability at least $1-2\delta$, it holds that*

$$\sum_{i=1}^{n}\mathbb{V}(P_{s,a,b,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i}) \leq n\overline{\nu}^{\mathrm{ref}} + 3H^2\sqrt{n\iota}. \tag{24}$$

Proof: Note that

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{V}(P_{s,a,b,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i}) &= \sum_{i=1}^{n}\left(P_{s,a,b,h}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i})^2 - (P_{s,a,b,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i})^2\right)\\ &= \sum_{i=1}^{n}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}))^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2 + \chi_3+\chi_4+\chi_5\\ &= n\overline{\nu}^{\mathrm{ref}} + \chi_3+\chi_4+\chi_5,\end{aligned} \tag{25}$$

where

$$\chi_3 = \sum_{i=1}^{n}\left((P_{s,a,b,h}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i})^2 - (\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}))^2\right),$$
$$\chi_4 = \frac{1}{n}\left(\sum_{i=1}^{n}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}P_{s,a,b,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2,$$
$$\chi_5 = \frac{1}{n}\left(\sum_{i=1}^{n}P_{s,a,b,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2 - \sum_{i=1}^{n}(P_{s,a,b,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i})^2.$$

By Azuma's inequality, with probability at least $1-\delta$ it holds that $|\chi_3| \leq H^2\sqrt{2n\iota}$.

By Azuma's inequality, with probability at least $1-\delta$, it holds that

$$|\chi_4| = \frac{1}{n}\left|\left(\sum_{i=1}^{n}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2 - \left(\sum_{i=1}^{n}P_{s,a,b,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2\right|$$

$$\leq 2H \left| \sum_{i=1}^{n} \overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) - \sum_{i=1}^{n} P_{s,a,b,h} \overline{V}_{h+1}^{\mathrm{ref},\ell_i} \right|$$
$$\leq 2H^2\sqrt{2n\iota}.$$

Moreover, $\chi_5 \leq 0$ by Cauchy-Schwartz inequality. Plugging the above inequalities gives the desired result. ■

Combining (22) with (24) gives

$$|\chi_1| \leq 2\sqrt{\frac{\underline{\nu}^{\mathrm{ref}}\iota}{n}} + \frac{5H\iota^{\frac{3}{4}}}{n^{\frac{3}{4}}} + \frac{2\sqrt{\iota}}{Tn} + \frac{2H\iota}{n}. \tag{26}$$

Similar to Lemma E.1, we have the following lemma.

Lemma E.2. *With probability at least $1-2\delta$, it holds that*

$$\sum_{i=1}^{\check{n}} \mathbb{V}(P_{s,a,b,h}, \overline{W}_{h+1}^{\mathrm{ref},\ell_i}) \leq \check{n}\check{\nu} + 3H^2\sqrt{\check{n}\iota}. \tag{27}$$

Combining (23) with (27) gives

$$|\chi_2| \leq 2\sqrt{\frac{\check{\nu}\iota}{\check{n}}} + \frac{5H\iota^{\frac{3}{4}}}{\check{n}^{\frac{3}{4}}} + \frac{2\sqrt{\iota}}{T\check{n}} + \frac{2H\iota}{\check{n}}. \tag{28}$$

Finally, combining (26) and (28), noting the definition of $\overline{\beta}$ with $(c_1, c_2, c_3) = (2, 2, 5)$, and taking a union bound over all probability events, we have that with probability at least $1 - 2(H^2T^3 + 3)\delta$, it holds that

$$\overline{\beta} \geq |\chi_1| + |\chi_2|. \tag{29}$$

which means $\overline{Q}_h^{k+1}(s,a,b) \geq Q_h^*(s,a,b)$.

Combining the two cases and taking the union bound over all steps, we have with probability at least $1 - T(2H^2T^3 + 7)\delta$, it holds that $\overline{Q}_h^{k+1}(s,a,b) \geq Q_h^*(s,a,b)$.

Next, we show that $V_h^*(s) \leq \overline{V}_h^{k+1}(s)$. Note that

$$\begin{aligned}
\overline{V}_h^{k+1}(s) &= (\mathbb{D}_{\pi_h^{k+1}} \overline{Q}_h^{k+1})(s) \\
&\geq \sup_{\mu \in \Delta_{\mathcal{A}}} (\mathbb{D}_{\mu \times \nu_h^{k+1}} \overline{Q}_h^{k+1})(s) && (30) \\
&\geq \sup_{\mu \in \Delta_{\mathcal{A}}} (\mathbb{D}_{\mu \times \nu_h^{k+1}} Q_h^*)(s) && (31) \\
&\geq \sup_{\mu \in \Delta_{\mathcal{A}}} \inf_{\nu \in \Delta_{\mathcal{B}}} (\mathbb{D}_{\mu \times \nu} Q_h^*)(s) \\
&= V_h^*(s),
\end{aligned}$$

where (30) follows from the property of the CCE oracle, (31) follows because $\overline{Q}_h^{k+1}(s,a,b) \geq \overline{Q}_h^*(s,a,b)$, which has just been proved.

Step ii: We show the inequalities for the pessimistic action-value function and value function below.

The two inequalities with respect to pessimistic (action-)value functions clearly hold for $k = 1$. For $k \geq 2$, suppose $Q_h^*(s,a,b) \geq \underline{Q}_h^u(s,a,b)$ and $V_h^*(s) \geq \underline{V}_h^u(s)$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $u \in [1,k]$. Now we fix tuple (s,a,b,h) and we only need to consider the case when $\underline{Q}_h(s,a,b)$ and $\underline{V}_h(s)$ are updated.

We show $Q_h^*(s,a,b) \geq \underline{Q}_h^{k+1}(s,a,b)$. Note that

$$\underline{Q}_h^{k+1}(s,a,b) \leftarrow \min\left\{ r_h(s,a,b) + \frac{\check{v}}{\check{n}} + \gamma, r_h(s,a,b) + \frac{\underline{\mu}^{\mathrm{ref}}}{n} + \frac{\check{\mu}}{\check{n}} + \underline{\beta}, \underline{Q}_h^k(s,a,b) \right\},$$

and we have two non-trivial cases.

For the first case, by Hoeffding's inequality, with probability at least $1-\delta$, it holds that

$$\begin{aligned}
\underline{Q}_h^{k+1}(s,a,b) &= r_h(s,a,b) + \frac{\check{\underline{v}}}{\check{n}} - \gamma \\
&= r_h(s,a,b) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \underline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - 2\sqrt{\frac{H^2}{\check{n}}\iota} \\
&\le r_h(s,a,b) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} V_{h+1}^{*}(s_{h+1}^{\check{\ell}_i}) - 2\sqrt{\frac{H^2}{\check{n}}\iota} \qquad (32)\\
&\le r_h(s,a,b) + (P_h V_{h+1}^*)(s,a,b) \qquad (33)\\
&= Q_h^*(s,a,b),
\end{aligned}$$

where (32) follows from the induction hypothesis $\underline{V}_{h+1}^{u}(s) \ge V^*(s)$ for all $u \in [k]$, and (33) follows from Azuma-Hoeffding's inequality.

For the second case, we have

$$\begin{aligned}
&\underline{Q}_h^{k+1}(s,a,b) = r_h(s,a,b) + \frac{\underline{\mu}^{\mathrm{ref}}}{n} + \frac{\check{\underline{\mu}}}{\check{n}} - \underline{\beta} \\
&= r_h(s,a,b) + \frac{1}{n}\sum_{i=1}^{n} \underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \left(\underline{V}_{h+1}^{\check{\ell}_i} - \underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)(s_{h+1}^{\check{\ell}_i}) - \underline{\beta} \\
&= r_h(s,a,b) + \left(P_h\left(\frac{1}{n}\sum_{i=1}^{n}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)\right)(s,a,b) + \left(P_h\left(\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(\underline{V}_{h+1}^{\check{\ell}_i} - \underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)\right)\right)(s,a,b) \\
&\qquad + \underline{\chi}_1 + \underline{\chi}_2 - \underline{\beta} \\
&\le r_h(s,a,b) + \left(P_h\left(\frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\underline{V}_{h+1}^{\check{\ell}_i}\right)\right)(s,a,b) + \underline{\chi}_1 + \underline{\chi}_2 - \underline{\beta} \qquad (34)\\
&\le r_h(s,a,b) + \left(P_h V_{h+1}^*\right)(s,a,b) + \underline{\chi}_1 + \underline{\chi}_2 - \underline{\beta} \qquad (35)\\
&= \underline{Q}_h^*(s,a,b) + \underline{\chi}_1 + \underline{\chi}_2 - \underline{\beta},
\end{aligned}$$

where

$$\begin{aligned}
\underline{\chi}_1(k,h) &= \frac{1}{n}\sum_{i=1}^{n}\left(\underline{V}_h^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) - \left(P_h\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)(s,a,b)\right), \\
\underline{W}_{h+1}^{\ell} &= \underline{V}_{h+1}^{\ell} - \underline{V}_{h+1}^{\mathrm{ref},\ell} \\
\underline{\chi}_2(k,h) &= \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(\underline{W}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \left(P_h\underline{W}_{h+1}^{\check{\ell}_i}\right)(s,a,b)\right).
\end{aligned}$$

Here, (34) follows from the fact that $\underline{V}_{h+1}^{\mathrm{ref},u}(s)$ is non-decreasing in u (since $\underline{V}_h^{\mathrm{ref}}(s)$ for a pair (s,h) can only be updated once and the updated value is obviously greater than the initial value 0), and (35) follows from the induction hypothesis $\underline{V}_{h+1}^{k}(s) \le V_{h+1}^*(s)$.

By Lemma I.2 with $\epsilon = \frac{1}{T^2}$, with probability at least $1 - 2(H^2T^3+1)\delta$ it holds

$$|\underline{\chi}_1(k,h)| \le 2\sqrt{\frac{\sum_{i=1}^{n}\mathbb{V}(P_{s,a,b,h}, \underline{V}_{h+1}^{\mathrm{ref},\ell_i})\iota}{n^2}} + \frac{2\sqrt{\iota}}{Tn} + \frac{2H\iota}{n}, \qquad (36)$$

$$|\underline{\chi}_2(k,h)| \le 2\sqrt{\frac{\sum_{i=1}^{\check{n}}\mathbb{V}(P_{s,a,b,h}, \underline{V}_{h+1}^{\mathrm{ref},\ell_i})\iota}{\check{n}^2}} + \frac{2\sqrt{\iota}}{T\check{n}} + \frac{2H\iota}{\check{n}}. \qquad (37)$$

Lemma E.3. *With probability at least* $1-2\delta$*, it holds that*

$$\sum_{i=1}^{n}\mathbb{V}(P_{s,a,b,h},\underline{V}_{h+1}^{\mathrm{ref},\ell_i})\le n\underline{\nu}^{\mathrm{ref}}+3H^2\sqrt{n\iota} \tag{38}$$

Proof: Note that

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{V}(P_{s,a,b,h},\underline{V}_{h+1}^{\mathrm{ref},\ell_i}) &= \sum_{i=1}^{n}\left(P_{s,a,b,h}(\underline{V}_{h+1}^{\mathrm{ref},\ell_i})^2-(P_{s,a,b,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i})^2\right)\\ &= \sum_{i=1}^{n}(\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}))^2-\frac{1}{n}\left(\sum_{i=1}^{n}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2+\underline{\chi}_3+\underline{\chi}_4+\underline{\chi}_5\\ &= n\underline{\nu}^{\mathrm{ref}}+\underline{\chi}_3+\underline{\chi}_4+\underline{\chi}_5,\end{aligned} \tag{39}$$

where

$$\begin{aligned}\underline{\chi}_3 &= \sum_{i=1}^{n}\left((P_{s,a,b,h}(\underline{V}_{h+1}^{\mathrm{ref},\ell_i})^2-(\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}))^2\right),\\ \underline{\chi}_4 &= \frac{1}{n}\left(\sum_{i=1}^{n}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2-\frac{1}{n}\left(\sum_{i=1}^{n}P_{s,a,b,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2,\\ \underline{\chi}_5 &= \frac{1}{n}\left(\sum_{i=1}^{n}P_{s,a,b,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2-\sum_{i=1}^{n}(P_{s,a,b,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i})^2.\end{aligned}$$

By Azuma's inequality, with probability at least $1-\delta$ it holds that $|\underline{\chi}_3|\le H^2\sqrt{2n\iota}$.

By Azuma's inequality, with probability at least $1-\delta$, it holds that

$$\begin{aligned}|\underline{\chi}_4| &= \frac{1}{n}\left|\left(\sum_{i=1}^{n}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2-\left(\sum_{i=1}^{n}P_{s,a,b,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2\right|\\ &\le 2H\left|\sum_{i=1}^{n}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})-\sum_{i=1}^{n}P_{s,a,b,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right|\\ &\le 2H^2\sqrt{2n\iota}.\end{aligned}$$

Moreover, $\chi_5\le 0$ by Cauchy-Schwartz inequality. Substituting the above inequalities gives the desired result. ■

Combining (36) with (38) gives

$$|\underline{\chi}_1|\le 2\sqrt{\frac{\underline{\nu}^{\mathrm{ref}}\iota}{n}}+\frac{5H\iota^{\frac{3}{4}}}{n^{\frac{3}{4}}}+\frac{2\sqrt{\iota}}{Tn}+\frac{2H\iota}{n}. \tag{40}$$

Similar to Lemma E.3, we have the following lemma.

Lemma E.4. *With probability at least* $1-2\delta$*, it holds that*

$$\sum_{i=1}^{\check{n}}\mathbb{V}(P_{s,a,b,h},\underline{W}_{h+1}^{\mathrm{ref},\ell_i})\le\check{n}\underline{\check{\nu}}+3H^2\sqrt{\check{n}\iota}. \tag{41}$$

Combining (37) with (41) gives

$$|\underline{\chi}_2|\le 2\sqrt{\frac{\underline{\check{\nu}}\iota}{\check{n}}}+\frac{5H\iota^{\frac{3}{4}}}{\check{n}^{\frac{3}{4}}}+\frac{2\sqrt{\iota}}{T\check{n}}+\frac{2H\iota}{\check{n}}. \tag{42}$$

Finally, combining (40) and (42), noting the definition of $\underline{\beta}$ with $(c_1, c_2, c_3) = (2, 2, 5)$, and taking a union bound over all probability events, we have that with probability at least $1 - 2(H^2T^3 + 3)\delta$, it holds that

$$\underline{\beta} \geq |\underline{\chi}_1| + |\underline{\chi}_2|. \tag{43}$$

which gives $\underline{Q}_h^{k+1}(s, a, b) \leq Q_h^*(s, a, b)$.

Combining the two cases and taking union bound over all steps, we have with probability at least $1 - T(2H^2T^3 + 7)\delta$, it holds that $\underline{Q}_h^{k+1}(s, a, b) \leq Q_h^*(s, a, b)$.

We show that $V_h^*(s) \leq \underline{V}_h^k(s)$. Note that

$$\begin{aligned}
\underline{V}_h^{k+1}(s) &= (\mathbb{D}_{\pi_h^{k+1}} \underline{Q}_h^{k+1})(s) \\
&\leq \inf_{\nu \in \Delta_{\mathcal{B}}} (\mathbb{D}_{\mu_h^{k+1} \times \nu} \underline{Q}_h^{k+1})(s) && (44) \\
&\leq \inf_{\nu \in \Delta_{\mathcal{B}}} (\mathbb{D}_{\mu_h^{k+1} \times \nu} Q_h^*)(s) && (45) \\
&\leq \inf_{\nu \in \Delta_{\mathcal{B}}} \sup_{\mu \in \Delta_{\mathcal{A}}} (\mathbb{D}_{\mu \times \nu} Q_h^*)(s) \\
&= V_h^*(s),
\end{aligned}$$

where (44) follows from the property of the CCE oracle, (45) follows because $\underline{Q}_h^{k+1}(s, a, b) \leq \underline{Q}_h^*(s, a, b)$, which has just been proved.

The entire proof is completed by combining **step i** and **step ii**, and taking a union bound over all probability events.

F. Proof of Lemma D.2 (Step II)

First, by Hoeffing's inequality, for any $(k, h) \in [K] \times [H]$, with probability at least $1 - 2T\delta$ it holds that

$$\left| \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \overline{Q}_h^k(s_h^k, a_h^k, b_h^k) \right| \leq \gamma_h^k, \qquad \left| \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \underline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \underline{Q}_h^k(s_h^k, a_h^k, b_h^k) \right| \leq \gamma_h^k.$$

The entire proof will be conditioned on the above event.

For any weight sequence $\{w_k\}_{k=1}^K$ where $w_k \geq 0$, let $\|w\|_\infty = \max_{1 \leq k \leq K} w_k$ and $\|w\|_1 = \sum_{k=1}^K w_k$.

By the update rule of the action-value function, we have

$$\begin{aligned}
\Delta_h^k &= (\overline{V}_h^k - \underline{V}_h^k)(s_h^k) \\
&= \zeta_h^k + (\overline{Q}_h^k - \underline{Q}_h^k)(s_h^k, a_h^k, b_h^k) \\
&\leq \zeta_h^k + 2\gamma_h^k + \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} (\overline{V}_{h+1}^{\check{\ell}_i} - \underline{V}_{h+1}^{\check{\ell}_i})(s_{h+1}^{\check{\ell}_i}) + H\mathbf{1}\{n_h^k = 0\} \\
&= \zeta_h^k + 2\gamma_h^k + \frac{1}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \Delta_{h+1}^{\check{\ell}_i} + H\mathbf{1}\{n_h^k = 0\}. && (46)
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{k=1}^K \frac{w_k}{\check{n}_h^k} \sum_{i=1}^{\check{n}_h^k} \Delta_{h+1}^{\check{\ell}_i} &= \sum_{j=1}^K \frac{w_j}{\check{n}_h^j} \sum_{i=1}^{\check{n}_h^j} \Delta_{h+1}^{\check{\ell}_{h,i}^j} \\
&= \sum_{j=1}^K \frac{w_j}{\check{n}_h^j} \sum_{k=1}^K \Delta_{h+1}^k \sum_{i=1}^{\check{n}_h^j} \mathbf{1}\{k = \check{\ell}_{h,i}^j\}
\end{aligned}$$

$$= \sum_{k=1}^{K} \Delta_{h+1}^k \sum_{j=1}^{K} \frac{w_j}{\check{n}_h^j} \sum_{i=1}^{\check{n}_h^j} \mathbf{1}\{k = \check{\ell}_{h,i}^j\}$$
$$= \sum_{k=1}^{K} \widetilde{w}_k \Delta_{h+1}^k, \tag{47}$$

where we define $\widetilde{w}_k = \sum_{j=1}^{K} \frac{w_j}{\check{n}_h^j} \sum_{i=1}^{\check{n}_h^j} \mathbf{1}\{k = \check{\ell}_{h,i}^j\}$. Similar to the proof of Lemma D.4, we have

$$\|\widetilde{w}\|_\infty = \max_k \widetilde{w}_k \le (1 + \frac{1}{H}) \|w\|_\infty . \tag{48}$$

Moreover,

$$\|\widetilde{w}\|_1 = \sum_{k=1}^{K} \sum_{j=1}^{K} \frac{w_j}{\check{n}_h^j} \sum_{i=1}^{\check{n}_h^j} \mathbf{1}\{k = \check{\ell}_{h,i}^j\} = \sum_{j=1}^{K} \frac{w_j}{\check{n}_h^j} \sum_{k=1}^{K} \sum_{i=1}^{\check{n}_h^j} \mathbf{1}\{k = \check{\ell}_{h,i}^j\} = \sum_{j=1}^{K} w_j = \|w\|_1 . \tag{49}$$

Combining (46), (47), (48) and (49), we have

$$\sum_{k=1}^{K} w_k \Delta_h^k \le \sum_{k=1}^{K} w_k \zeta_h^k + 2 \sum_{k=1}^{K} w_k \gamma_h^k + \sum_{k=1}^{K} \widetilde{w}_k \Delta_{h+1}^k + H \sum_{k=1}^{K} w_k \mathbf{1}\{n_h^k = 0\}$$
$$\le \sum_{k=1}^{K} w_k \zeta_h^k + 2 \sum_{k=1}^{K} w_k \gamma_h^k + \sum_{k=1}^{K} \widetilde{w}_k \Delta_{h+1}^k + SABH^2 \|w\|_\infty . \tag{50}$$

By Azuma-Hoeffding's inequality, with probability at least $1 - H\delta$, it holds that for any $h \in [H]$

$$\sum_{k=1}^{K} w_k \zeta_h^k \le \sqrt{2} H \iota \sqrt{\sum_{k=1}^{K} w_k} \le \sqrt{2} H \iota \|w\|_\infty . \tag{51}$$

We now bound the second term of (50). Define $\Xi(s,a,b,j) = \sum_{k=1}^{K} w_k \mathbf{1}\{\check{n}_h^k = e_j, (s_h^k, a_h^k, b_h^k) = (s,a,b)\}$ and $\Xi(s,a,b) = \sum_{j \ge 1} \Xi(s,a,b,j)$. Similar to (48) and (49), we then have $\Xi(s,a,b,j) \le \|w\|_\infty (1 + \frac{1}{H}) e_j$ and $\sum_{s,a} \Xi(s,a,b) = \sum_k w_k$. Then

$$\sum_k w_k \gamma_h^k = \sum_k 2\sqrt{H^2 \iota} w_k \sqrt{\frac{1}{\check{n}_h^k}}$$
$$= 2\sqrt{H^2 \iota} \sum_{s,a,b,j} \sqrt{\frac{1}{e_j}} \sum_{j=1}^{K} w_k \mathbf{1}\{\check{n}_h^k = e_j, (s_h^k, a_h^k, b_h^k) = (s,a,b)\}$$
$$= 2\sqrt{H^2 \iota} \sum_{s,a,b} \sum_{j \ge 1} \Xi(s,a,b,j) \sqrt{\frac{1}{e_j}}.$$

Fix (s,a,b) and consider $\sum_{j \ge 1} \Xi(s,a,b,j) \sqrt{\frac{1}{e_j}}$. Note that $\sqrt{\frac{1}{e_j}}$ is decreasing in j. Given $\sum_{j \ge 1} \Xi(s,a,b,j) = \Xi(s,a,b)$ is fixed, rearranging the inequality gives

$$\sum_{j \ge 1} \Xi(s,a,b,j) \sqrt{\frac{1}{e_j}} \le \sum_{j \ge 1} \sqrt{\frac{1}{e_j}} \|w\|_\infty (1 + \frac{1}{H}) e_j \mathbf{1} \left\{ \sum_{i=1}^{j-1} \|w\|_\infty (1 + \frac{1}{H}) e_i \le \Xi(s,a,b) \right\}$$
$$= \|w\|_\infty (1 + \frac{1}{H}) \sum_j \sqrt{e_j} \mathbf{1} \left\{ \sum_{i=1}^{j-1} \|w\|_\infty e_i \le \Xi(s,a,b) \right\}$$

$$\leq 10(1+\frac{1}{H})\sqrt{\|w\|_\infty H\Xi(s,a,b)}.$$

Therefore, by Cauchy-Schwartz inequality, we have

$$\begin{aligned}\sum_{k=1}^{K} w_k\gamma_h^k &\leq 2\sqrt{H^2\iota}\sum_{s,a,b} 10(1+\frac{1}{H})\sqrt{\|w\|_\infty H}\sqrt{\Xi(s,a,b)}\\ &\leq 20\sqrt{H^2\iota}(1+\frac{1}{H})\sqrt{\|w\|_\infty SABH\|w\|_1}. \end{aligned} \tag{52}$$

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

$$\sum_{k=1}^{K} w_k\Delta_h^k \leq \sum_{k=1}^{K}\widetilde{w}_k\Delta_{h+1}^k + (\sqrt{2}H\iota + SABH^2)\|w\|_\infty + 80H\sqrt{\|w\|_\infty SABH\|w\|_1\iota}. \tag{53}$$

We expand the expression by iterating over step $h+1,\cdots,H$,

$$\begin{aligned}\sum_{k=1}^{K} w_k\Delta_h^k &\leq (1+\frac{1}{H})^H\cdot H\cdot\left((\sqrt{2}H\iota + SABH^2)\|w\|_\infty + 80H\sqrt{\|w\|_\infty SABH\|w\|_1\iota}\right)\\ &\leq 6(H^2\iota + SABH^3)\|w\|_\infty + 240H^{\frac{5}{2}}\sqrt{\|w\|_\infty SAB\|w\|_1\iota}.\end{aligned}$$

Now we set $w_k = \mathbf{1}\{\Delta_h^k \geq \epsilon\}$, and obtain

$$\sum_{k=1}^{K}\mathbf{1}\{\Delta_h^k\geq\epsilon\}\Delta_h^k \leq 6(H^2\iota+SABH^3)\|w\|_\infty + 240H^{\frac{5}{2}}\sqrt{\|w\|_\infty SAB\iota\sum_{k=1}^{K}\mathbf{1}\{\Delta_h^k\geq\epsilon\}}.$$

Note that $\|w\|_\infty$ is either 0 or 1. If $\|w\|_\infty = 0$, the claim obviously holds. In the case when $\|w\|_\infty = 1$, solving the following quadratic equation (ignoring coefficients) with respect to $\left(\sum_{k=1}^{K}\mathbf{1}\{\Delta_h^k\geq\epsilon\}\right)^{1/2}$ gives the desired result

$$\epsilon\left(\sum_{k=1}^{K}\mathbf{1}\{\Delta_h^k\geq\epsilon\}\right) - H^{5/2}(SAB\iota)^{1/2}\left(\sum_{k=1}^{K}\mathbf{1}\{\Delta_h^k\geq\epsilon\}\right)^{1/2} - (SABH^3+H^2\iota)\leq 0.$$

G. Proof of Lemma D.5 (Step IV)

The entire proof is conditioned on the successful events of Lemma D.1 and Lemma D.2, which occur with probability at least $1-O(H^2T^4)\delta$.

By the definition of Λ_{h+1}^k, we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\Lambda_{h+1}^k &= \underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\psi_{h+1}^k}_{T_1} + \underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\xi_{h+1}^k}_{T_2}\\ &\quad + 2\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\overline{\beta}_h^k}_{T_3} + 2\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\underline{\beta}_h^k}_{T_4}.\end{aligned} \tag{54}$$

We next bound each of the above four terms in one subsection, and summarize the final result in Appendix G.5.

G.1. Bound T_1

Recall the definition $\lambda_h^k(s) = \mathbf{1}\left\{n_h^k(s) < N_0\right\}$. Since ψ is always non-negative, we have

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\psi_{h+1}^k \\
&\le 3\sum_{h=1}^{H}\sum_{k=1}^{K}\psi_{h+1}^k \\
&= 3\sum_{h=1}^{H}\sum_{k=1}^{K}P_{s_h^k,a_h^k,b_h^k,h}\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(\overline{V}_{h+1}^{\mathrm{ref},\ell_i}-\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)-\left(\overline{V}_{h+1}^{\mathrm{REF}}-\underline{V}_{h+1}^{\mathrm{REF}}\right)\right) \\
&\le 3H\sum_{h=1}^{H}\sum_{k=1}^{K}P_{s_h^k,a_h^k,b_h^k,h}\left(\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\lambda_{h+1}^{\ell_i}\right) \\
&\le 3H\sum_{h=1}^{H}\sum_{j=1}^{K}\sum_{k=1}^{K}P_{s_h^k,a_h^k,b_h^k,h}\lambda_{h+1}^j\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbf{1}\{j=\ell_{h,i}^k\} \\
&\le 3H\sum_{h=1}^{H}\sum_{j=1}^{K}P_{s_h^j,a_h^j,b_h^j,h}\lambda_{h+1}^j\sum_{k=1}^{K}\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbf{1}\{j=\ell_{h,i}^k\} \qquad (55)\\
&\le 6(\log T+1)H\sum_{h=1}^{H}\sum_{k=1}^{K}P_{s_h^k,a_h^k,b_h^k,h}\lambda_{h+1}^k \qquad (56)\\
&\le 6(\log T+1)H\left(\sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_{h+1}^k(s_{h+1}^k)+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^k}\right)\lambda_{h+1}^k\right) \\
&\le 6(\log T+1)H\left(HSN_0+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^k}\right)\lambda_{h+1}^k\right) \\
&\le 6(\log T+1)H\left(HSN_0+2\sqrt{T\iota}\right), \qquad (57)
\end{aligned}
$$

where (55) follows from the fact that $\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbf{1}\{j=\ell_{h,i}^k\}\neq 0$ only if $(s_h^k,a_h^k,b_h^k)=(s_h^j,a_h^j,b_h^j)$, (56) follows because

$$
\sum_{k=1}^{K}\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbf{1}\{j=\ell_{h,i}^k\}\le\sum_{z:j\le\sum_{i=1}^{z-1}e_i\le T}\frac{e_z}{\sum_{i=1}^{z-1}e_i}\le 2(\log T+1),
$$

and (57) holds with probability at least $1-\delta$ by Azuma's inequality.

To conclude, with probability at least $1-\delta$, it holds that

$$
\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\psi_{h+1}^k\le O(\log T)\cdot(H^2SN_0+H\sqrt{T\iota}). \qquad (58)
$$

G.2. Term T_2

We first derive

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\xi_{h+1}^k \\
&=\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^{\check{\ell}_i}}\right)\left(\overline{V}_{h+1}^{\check{\ell}_i}-\underline{V}_{h+1}^{\check{\ell}_i}\right)
\end{aligned}
$$

$$= \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{j=1}^{K}(1+\frac{1}{H})^{h-1}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^j}\right)\left(\overline{V}_{h+1}^j-\underline{V}_{h+1}^j\right)\mathbf{1}\{\check{\ell}_{h,i}^k=j\}.$$

Note that $\check{\ell}_{h,i}^k = j$ if and only if $(s_h^k, a_h^k, b_h^k) = (s_h^j, a_h^j, b_h^j)$. Therefore,

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\xi_{h+1}^k\\
&\le \sum_{h=1}^{H}\sum_{j=1}^{K}(1+\frac{1}{H})^{h-1}\left(P_{s_h^j,a_h^j,b_h^j,h}-\mathbf{1}_{s_{h+1}^j}\right)\left(\overline{V}_{h+1}^j-\underline{V}_{h+1}^j\right)\sum_{k=1}^{K}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\mathbf{1}\{\check{\ell}_{h,i}^k=j\}\\
&= \sum_{h=1}^{H}\sum_{k=1}^{K}\theta_{h+1}^j\left(P_{s_h^j,a_h^j,b_h^j,h}-\mathbf{1}_{s_{h+1}^j}\right)\left(\overline{V}_{h+1}^j-\underline{V}_{h+1}^j\right),
\end{aligned}$$

where in the last equation we define $\theta_{h+1}^j = (1+\frac{1}{H})^{h-1}\sum_{k=1}^{K}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\mathbf{1}\{\check{\ell}_{h,i}^k=j\}$.

For $(j,h)\in[K]\times[H]$, let x_h^j be the number of elements in current state with respect to (s_h^j, a_h^j, b_h^j, h) and $\tilde{\theta}_{h+1}^j := (1+\frac{1}{H})^{h-1}\frac{\lfloor(1+\frac{1}{H})x_h^j\rfloor}{x_h^j}\le 3$. Define $\mathcal{K}=\{(k,h):\theta_{h+1}^k=\tilde{\theta}_{h+1}^k\}$. Note that if k is before the second last stage of the tuple (s_h^k, a_h^k, b_h^k, h), then we have that $\theta_{h+1}^k=\tilde{\theta}_{h+1}^k$ and $(k,h)\in\mathcal{K}$. Given $(k,h)\in\mathcal{K}$, s_{h+1}^k follows the transition $P_{s_h^k,a_h^k,b_h^k,h}$.

Let $\mathcal{K}_h^{\perp}(s,a,b)=\{k:(s_h^k,a_h^k,b_h^k)=(s,a,b)$, where k is in the second last stage of $(s,a,b,h)\}$. Note that for different j,k, if $(s_h^k,a_h^k,b_h^k)=(s_h^j,a_h^j,b_h^j)$ and j,k are in the same stage of (s_h^k,a_h^k,b_h^k,h), then $\theta_{h+1}^k=\theta_{h+1}^j$ and $\tilde{\theta}_{h+1}^k=\tilde{\theta}_{h+1}^j$. Denote θ_{h+1} and $\tilde{\theta}_{h+1}$ as $\theta_{h+1}(s,a,b)$ and $\tilde{\theta}_{h+1}(s,a,b)$ respectively for some $k\in\mathcal{K}_h^{\perp}(s,a,b)$.

We have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\xi_{h+1}^k\\
&= \sum_{(k,h)}\tilde{\theta}_{h+1}^k\left(P_{s_h^j,a_h^j,b_h^j,h}-\mathbf{1}_{s_{h+1}^j}\right)\left(\overline{V}_{h+1}^j-\underline{V}_{h+1}^j\right)\\
&\quad+\sum_{(k,h)}(\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\left(P_{s_h^j,a_h^j,b_h^j,h}-\mathbf{1}_{s_{h+1}^j}\right)\left(\overline{V}_{h+1}^j-\underline{V}_{h+1}^j\right)\\
&= \sum_{(k,h)}\tilde{\theta}_{h+1}^k\left(P_{s_h^j,a_h^j,b_h^j,h}-\mathbf{1}_{s_{h+1}^j}\right)\left(\overline{V}_{h+1}^j-\underline{V}_{h+1}^j\right)\\
&\quad+\sum_{(k,h)\in\overline{\mathcal{K}}}(\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\left(P_{s_h^j,a_h^j,b_h^j,h}-\mathbf{1}_{s_{h+1}^j}\right)\left(\overline{V}_{h+1}^j-\underline{V}_{h+1}^j\right).
\end{aligned}\tag{59}$$

Since $\tilde{\theta}_{h+1}^k$ is independent of s_{h+1}^k, by Azuma's inequality, with probability at least $1-\delta$, it holds that

$$\sum_{(k,h)}\tilde{\theta}_{h+1}^k\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^k}\right)\left(\overline{V}_{h+1}^k-\underline{V}_{h+1}^k\right)\le 6\sqrt{TH^2\iota}.\tag{60}$$

Moreover, we have

$$\begin{aligned}
&\sum_{(k,h)\in\overline{\mathcal{K}}}(\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^k}\right)\left(\overline{V}_{h+1}^k-\underline{V}_{h+1}^k\right)\\
&=\sum_{s,a,b,h}\sum_{(k,h)\in\overline{\mathcal{K}}}\mathbf{1}\{(s_h^k,a_h^k,b_h^k)=(s,a,b)\}(\theta_{h+1}^k-\tilde{\theta}_{h+1}^k)\left(P_{s_h^k,a_h^k,b_h^k,h}-\mathbf{1}_{s_{h+1}^k}\right)\left(\overline{V}_{h+1}^k-\underline{V}_{h+1}^k\right)
\end{aligned}$$

$$= \sum_{s,a,b,h} (\theta_{h+1}(s,a,b) - \tilde{\theta}_{h+1}(s,a)) \sum_{(k,h)\in\mathcal{K}_h^\perp(s,a)} (\theta_{h+1}^k - \tilde{\theta}_{h+1}^k) \left(P_{s_h^k,a_h^k,b_h^k,h} - \mathbf{1}_{s_{h+1}^k}\right) \left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)$$

$$\leq \sum_{s,a,b,h} \mathcal{O}(H)\sqrt{|\mathcal{K}_h^\perp(s,a,b)|\iota} \tag{61}$$

$$\leq \sum_{s,a,b,h} \mathcal{O}(H)\sqrt{\check{N}_h^{K+1}(s,a,b)\iota}$$

$$\leq \mathcal{O}(H)\sqrt{SABH\iota \sum_{s,a,b,h} \check{N}_h^{K+1}(s,a,b)} \tag{62}$$

$$\leq \mathcal{O}(H)\sqrt{SABH\iota(T/H)}, \tag{63}$$

where (61) holds with probability at least $1 - T\delta$ by Azuma's inequality and a union bound over all steps in $\overline{\mathcal{K}}$, (62) follows from Cauchy-Schwartz inequality, and (63) follows from the fact that the length of the last two stages for each (s,a,b,h) tuple is only $O(1/H)$ fraction of the total number of visits.

Combining (59), (60) and (63), we obtain that with probability at least $1-(T+1)\delta$, it holds that

$$\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\xi_{h+1}^k \leq \mathcal{O}(\sqrt{H^2SABT\iota}). \tag{64}$$

G.3. Term T_3

Note that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\overline{\beta}_h^k \\
&\leq 3\sum_{h=1}^{H}\sum_{k=1}^{K}\left(c_1\sqrt{\frac{\overline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota} + c_2\sqrt{\frac{\check{\overline{\nu}}_h^k}{\check{n}_h^k}\iota} + c_3\left(\frac{H\iota}{n_h^k} + \frac{H\iota}{\check{n}_h^k} + \frac{H\iota^{\frac{3}{4}}}{(n_h^k)^{\frac{3}{4}}} + \frac{H\iota^{\frac{3}{4}}}{(\check{n}_h^k)^{\frac{3}{4}}}\right)\right) \\
&\leq O\left(\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\sqrt{\frac{\overline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota} + \sqrt{\frac{\check{\overline{\nu}}_h^k}{\check{n}_h^k}\iota}\right)\right) + O(SABH^3\iota\log T + (SAB\iota)^{\frac{3}{4}}H^{\frac{5}{2}}T^{\frac{1}{4}}),
\end{aligned} \tag{65}$$

where (65) follows from Lemma I.3 with $\alpha = \frac{3}{4}$ and $\alpha = 1$.

Step i: We bound $\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\overline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota}$. We begin with the following technical lemmas.

Lemma G.1. *With probability at least $1-2T\delta$, it holds that for all s,a,b,h,k,*

$$\begin{aligned}
\overline{Q}_h^k(s,a,b) &\leq Q_h^{\pi^k}(s,a,b) + (H-h)\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right), \\
\underline{Q}_h^k(s,a,b) &\geq Q_h^{\pi^k}(s,a,b) - (H-h)\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right), \\
\overline{V}_h^k(s) &\leq V_h^{\pi^k}(s) + (H-h)\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right), \\
\underline{V}_h^k(s) &\geq V_h^{\pi^k}(s) - (H-h)\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right).
\end{aligned}$$

The proof is provided in Appendix G.3.1.

Lemma G.2. *Conditioned on the successful event of Lemma G.1, with probability at least $1-4\delta$, it holds that*

$$\overline{\nu}_h^{\mathrm{ref},k} - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, V_{h+1}^{\pi_k}) \leq 4H\beta + \frac{12H^2\beta + 18H^3SN_0}{n_h^k} + 20H^2\sqrt{\frac{\iota}{n_h^k}}.$$

The proof is provided in Appendix G.3.2.

Lemma G.3 (Lemma C.5 in [Jin et al., 2018]). *With probability at least $1-\delta$, it holds that*

$$\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, V_{h+1}^{\pi^k}) \le \mathcal{O}(HT + H^3\iota).$$

Combining Lemma G.2, Lemma G.3 and Lemma I.3 (see Appendix I), we have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\overline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota} \\
&\le \sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, V_{h+1}^{\pi^k})}{n_h^k}\iota} \\
&\quad + \sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\left(\frac{4H\beta}{n_h^k} + \frac{12H^2\beta + 18H^3SN_0}{(n_h^k)^2} + 20H^2\frac{\iota^{\frac{1}{2}}}{(n_h^k)^{\frac{3}{2}}}\right)\iota} \\
&\le O\left(\sum_{s,a,b,h}\sqrt{N_h^{K+1}(s,a,b)\mathbb{V}(P_{s,a,b,h}, V_{h+1}^{\pi^k})\iota}\right) \\
&\quad + O\left(\sum_{s,a,b,h}\sqrt{N_h^{K+1}(s,a,b)H\beta\iota} + (S^{\frac{3}{2}}ABH^{\frac{5}{2}}N_0^{\frac{1}{2}} + SABH^2\beta^{\frac{1}{2}})\iota^{\frac{1}{2}}\log T + (SAB\iota)^{\frac{3}{4}}H^{\frac{7}{4}}T^{\frac{1}{4}}\right) \\
&\le O\left(\sqrt{SABH^2T\iota} + \sqrt{SABH^2\beta T\iota} + (S^{\frac{3}{2}}ABH^{\frac{5}{2}}N_0^{\frac{1}{2}} + SABH^2\beta^{\frac{1}{2}})\iota^{\frac{1}{2}}\log T + (SAB\iota)^{\frac{3}{4}}H^{\frac{7}{4}}T^{\frac{1}{4}}\right). \qquad (66)
\end{aligned}$$

Step ii: We bound $\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\check{\nu}_h^k}{\check{n}_h^k}\iota}$.

By Lemma D.1, Lemma D.2 and Corollary D.3, we have

$$\begin{aligned}
\check{\nu}_h^k &\le \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\check{\ell}_i} - \overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)^2(s_{h+1}^{\check{\ell}_i}) \\
&\le \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\check{\ell}_i} - \underline{V}_{h+1}^{\check{\ell}_i}\right)^2(s_{h+1}^{\check{\ell}_i}) + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i} - \underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)^2(s_{h+1}^{\check{\ell}_i}) \\
&\le \frac{2}{\check{n}_h^k}H^2SN_0 + 2\beta^2.
\end{aligned}$$

Combining the above inequality with Lemma I.3, we obtain

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\check{\nu}_h^k\iota}{\check{n}_h^k}} \le \mathcal{O}\left(\sqrt{SABH^3\beta^2T\iota} + SABH^3\sqrt{SN_0\iota}\log T\right). \qquad (67)$$

Combining (65), (66) and (67), we obtain that with probability at least $1 - O(T)\delta$, it holds that

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\overline{\beta}_h^k \le O\big(&\sqrt{SABH^2T\iota} + \sqrt{SABH^2\beta T\iota} + \sqrt{SABH^3\beta^2T\iota} \\
&+ S^{\frac{3}{2}}ABH^3N_0^{\frac{1}{2}}\iota\log T + SABH^2\beta^{\frac{1}{2}}\iota^{\frac{1}{2}}\log T + (SAB\iota)^{\frac{3}{4}}H^{\frac{5}{2}}T^{\frac{1}{4}}\big). \qquad (68)
\end{aligned}$$

G.3.1. Proof of Lemma G.1

Fix an episode k. The proof is based on induction over $h = H, H-1, \ldots, 1$. Note first that the claim clearly holds for $h = H$. Assume the inequalities hold at step $h+1$.

By the update rule of the action-value function, we have

$$\begin{aligned}
\overline{Q}_h^k(s,a,b) &\le r_h(s,a,b) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) + \gamma \\
&= r_h(s,a,b) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \overline{V}_{h+1}^{k}(s_{h+1}^{\check{\ell}_i}) + \gamma + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \left(\overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \overline{V}_{h+1}^{k}(s_{h+1}^{\check{\ell}_i})\right) \\
&\le r_h(s,a,b) + P_{s,a,b,h}\overline{V}_{h+1}^{k} + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \left(\overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \overline{V}_{h+1}^{k}(s_{h+1}^{\check{\ell}_i})\right) \qquad (69) \\
&\le r_h(s,a,b) + P_{s,a,b,h}V_{h+1}^{\pi^k} + (H-h+1)\left(\beta + \frac{HSN_0}{\check{n}}\right) \\
&\qquad + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \left(\overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \overline{V}_{h+1}^{k}(s_{h+1}^{\check{\ell}_i})\right) \qquad (70) \\
&\le Q_h^{\pi^k} + (H-h+1)\left(\beta + \frac{HSN_0}{\check{n}}\right) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} \left(\overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \underline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i})\right) \qquad (71) \\
&\le Q_h^{\pi^k}(s,a,b) + (H-h+1)\left(\beta + \frac{HSN_0}{\check{n}}\right) + \frac{1}{\check{n}}\sum_{i=1}^{\check{n}} (H\lambda_{h+1}^{\check{\ell}_i} + \beta) \\
&\le Q_h^{\pi^k}(s,a,b) + (H-h)\left(\beta + \frac{HSN_0}{\check{n}}\right), \qquad (72)
\end{aligned}$$

where (69) holds with probability at least $1-\delta$ by Azuma's inequality, (70) follows from the induction hypothesis, and (71) follows from Lemma D.1.

Moreover, by the update rule of the value function, we have

$$\begin{aligned}
\overline{V}_h^k(s) &= \mathbb{E}_{(a,b)\sim\pi_k}\overline{Q}_h^k(s,a,b) \\
&\le \mathbb{E}_{(a,b)\sim\pi_k} Q_h^{\pi^k}(s,a,b) + (H-h)\left(\beta + \frac{HSN_0}{\check{n}}\right) \\
&\le V_h^{\pi^k}(s) + (H-h)\left(\beta + \frac{HSN_0}{\check{n}}\right).
\end{aligned}$$

The other direction for the pessimistic (action-)value function can be proved similarly. Finally, taking the union bound over all steps gives the desired result.

G.3.2. Proof of Lemma G.2

We first provide bound on $\overline{\nu}_h^{\mathrm{ref},k} - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, \overline{V}_{h+1}^{\mathrm{ref},\ell_i})$. Recall (25) that

$$\overline{\nu}^{\mathrm{ref}} - \frac{1}{n_h^k}\sum_{i=1}^{n_h^k} \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, \overline{V}_{h+1}^{\mathrm{ref},\ell_i}) = -\frac{1}{n_h^k}(\chi_6 + \chi_7 + \chi_8),$$

where

$$\begin{aligned}
\chi_6 &= \sum_{i=1}^{n_h^k}\left((P_{s_h^k,a_h^k,b_h^k,h}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i})^2 - (\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}))^2\right), \\
\chi_7 &= \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2 - \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2,
\end{aligned}$$

$$\chi_8 = \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2 - \sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i})^2.$$

By Azuma's inequality, with probability at least $1-2\delta$, it holds that

$$|\chi_6| \le H^2\sqrt{2n_h^k\iota}, \quad |\chi_7| \le 2H^2\sqrt{2n_h^k\iota}.$$

Moreover, we have

$$\begin{aligned}
-\chi_8 &= \sum_{i=1}^{n_h^k}\left(P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2 - \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2 \\
&\le \sum_{i=1}^{n_h^k}\left(P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2 - \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{REF}}\right)^2 \qquad (73)\\
&= \sum_{i=1}^{n_h^k}\left(\left(P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2 - \left(P_{s_h^k,a_h^k,b_h^k,h}\overline{V}_{h+1}^{\mathrm{REF}}\right)^2\right) \\
&\le 2H^2\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\lambda_{h+1}^{\ell_i} \\
&= 2H^2\left(\sum_{i=1}^{n_h^k}\lambda_{h+1}^{\ell_i}(s_{h+1}^{\ell_i}) + \sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,b_h^k,h} - \mathbf{1}_{s_{h+1}^{\ell_i}})\lambda_{h+1}^{\ell_i}\right) \\
&\le 2H^2SN_0 + 2H^2\sqrt{2n_h^k\iota}, \qquad (74)
\end{aligned}$$

where (73) follows from the fact that $\overline{V}_{h+1}^{\mathrm{ref},k} \ge \overline{V}_{h+1}^{\mathrm{REF}}$ for any k,h, and (74) holds with probability at least $1-\delta$ by Azuma's inequality.

We have

$$\overline{\nu}_h^{\mathrm{ref},k} - \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i}) \le \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}}. \qquad (75)$$

Therefore,

$$\begin{aligned}
&\overline{\nu}_h^{\mathrm{ref},k} - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},V_{h+1}^{\pi^k}) \\
&= \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i}) - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},V_{h+1}^{\pi^k})\right) + \left(\overline{\nu}_h^{\mathrm{ref},k} - \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i})\right) \\
&\le \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},\overline{V}_{h+1}^{\mathrm{ref},\ell_i}) - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},V_{h+1}^{\pi^k})\right) + \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}} \qquad (76)\\
&\le \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}\left|P_{s_h^k,a_h^k,b_h^k,h}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i} - V_{h+1}^{\pi^k})\right| + \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}} \\
&= \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}\left|P_{s_h^k,a_h^k,b_h^k,h}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i} - V_{h+1}^{\pi^k} + V_{h+1}^* - V_{h+1}^*) - H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right) + H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)\right| \\
&\qquad + \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}}
\end{aligned}$$

$$\leq \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i} - V_{h+1}^*) + \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\left(V_{h+1}^{\pi^k} - V_{h+1}^* + H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)\right)$$
$$+ \frac{4H^2}{n_h^k}\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right) + \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}} \tag{77}$$

$$\leq \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}(\overline{V}_{h+1}^{\mathrm{ref},\ell_i} - V_{h+1}^*)(s_{h+1}^{\ell_i}) + \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}\left((V_{h+1}^{\pi^k} - V_{h+1}^*)(s_{h+1}^{\ell_i}) + H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)\right)$$
$$+ \frac{4H^2}{n_h^k}\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right) + \frac{2H^2SN_0}{n_h^k} + 20H^2\sqrt{\frac{\iota}{n_h^k}} \tag{78}$$

$$\leq \left(4H\beta + \frac{4H^2SN_0}{n_h^k}\right) + \frac{8H^2}{n_h^k}\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)$$
$$+ \frac{4H^2}{n_h^k}\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right) + \frac{2H^2SN_0}{n_h^k} + 20H^2\sqrt{\frac{\iota}{n_h^k}} \tag{79}$$

$$= 4H\beta + \frac{12H^2\beta}{n_h^k} + 20H^2\sqrt{\frac{\iota}{n_h^k}} + \frac{6H^2SN_0}{n_h^k} + \frac{12H^3SN_0}{n_h^k\check{n}_h^k}$$
$$\leq 4H\beta + \frac{12H^2\beta + 18H^3SN_0}{n_h^k} + 20H^2\sqrt{\frac{\iota}{n_h^k}},$$

where (76) follows from (75), (77) follows from Lemma D.1 and Lemma G.1, (78) holds with probability at least $1-2\delta$ by Azuma's inequality, and (79) follows from Lemma D.1 and Lemma G.1.

G.4. Term T_4

The proof is similar to that for the term $\sum_{h=1}^H\sum_{k=1}^K(1+\frac{1}{H})^{h-1}\overline{\beta}_h^k$. In the following, we will present the key steps, and provide the proof whenever necessary.

By Lemma I.3, we have

$$\sum_{h=1}^H\sum_{k=1}^K(1+\frac{1}{H})^{h-1}\overline{\beta}_h^k$$
$$\leq 3\sum_{h=1}^H\sum_{k=1}^K\left(c_1\sqrt{\frac{\overline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota} + c_2\sqrt{\frac{\check{\overline{\nu}}_h^k}{\check{n}_h^k}\iota} + c_3\left(\frac{H\iota}{n_h^k} + \frac{H\iota}{\check{n}_h^k} + \frac{H\iota^{\frac{3}{4}}}{(n_h^k)^{\frac{3}{4}}} + \frac{H\iota^{\frac{3}{4}}}{(\check{n}_h^k)^{\frac{3}{4}}}\right)\right)$$
$$\leq O\left(\sum_{h=1}^H\sum_{k=1}^K\left(\sqrt{\frac{\overline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota} + \sqrt{\frac{\check{\overline{\nu}}_h^k}{\check{n}_h^k}\iota}\right)\right) + O(SABH^3\iota\log T + (SAB\iota)^{\frac{3}{4}}H^{\frac{5}{2}}T^{\frac{1}{4}}). \tag{80}$$

Step i: Bound term $\sum_{h=1}^H\sum_{k=1}^K\sqrt{\frac{\underline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota}$.

Lemma G.4. *Conditioned on the successful event of Lemma G.1, with probability at least $1-4\delta$, it holds that*

$$\underline{\nu}_h^{\mathrm{ref},k} - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, V_{h+1}^{\pi_k}) \leq 4H\beta + \frac{12H^2\beta + 18H^3SN_0}{n_h^k} + 20H^2\sqrt{\frac{\iota}{n_h^k}}.$$

The proof is provided in Appendix G.4.1.

Combining Lemma G.3, Lemma G.4 and Lemma I.3, we have

$$\sum_{h=1}^H\sum_{k=1}^K\sqrt{\frac{\underline{\nu}_h^{\mathrm{ref},k}}{n_h^k}\iota}$$

$$\leq O\left(\sqrt{SABH^2T\iota}+\sqrt{SABH^2\beta T\iota}+(S^{\frac{3}{2}}ABH^{\frac{5}{2}}N_0^{\frac{1}{2}}+SABH^2\beta^{\frac{1}{2}})\iota^{\frac{1}{2}}\log T+(SAB\iota)^{\frac{3}{4}}H^{\frac{7}{4}}T^{\frac{1}{4}}\right). \tag{81}$$

Step ii: Bound $\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\check{\underline{\nu}}_h^k}{\check{n}_h^k}\iota}$. By Lemma D.1, Lemma D.2 and Corollary D.3, we have

$$\begin{aligned}
\check{\underline{\nu}}_h^k &\leq \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\underline{V}_{h+1}^{\check{\ell}_i}-\underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)^2(s_{h+1}^{\check{\ell}_i})\\
&\leq \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\check{\ell}_i}-\underline{V}_{h+1}^{\check{\ell}_i}\right)^2(s_{h+1}^{\check{\ell}_i})+\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}-\underline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)^2(s_{h+1}^{\check{\ell}_i})\\
&\leq \frac{2}{\check{n}_h^k}H^2SN_0+2\beta^2.
\end{aligned}$$

Combining the above inequality with Lemma I.3, we obtain

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\check{\underline{\nu}}_h^k\iota}{\check{n}_h^k}}\leq\mathcal{O}\left(\sqrt{SABH^3\beta^2T\iota}+SABH^3\sqrt{SN_0\iota}\log T\right). \tag{82}$$

Therefore, combining (80), (81) and (82) gives that with probability at least $1-O(T)\delta$, it holds that

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\overline{\beta}_h^k \leq O\big(&\sqrt{SABH^2T\iota}+\sqrt{SABH^2\beta T\iota}+\sqrt{SABH^3\beta^2T\iota}\\
&+S^{\frac{3}{2}}ABH^3N_0^{\frac{1}{2}}\iota\log T+SABH^2\beta^{\frac{1}{2}}\iota^{\frac{1}{2}}\log T+(SAB\iota)^{\frac{3}{4}}H^{\frac{5}{2}}T^{\frac{1}{4}}\big).
\end{aligned} \tag{83}$$

G.4.1. Proof of Lemma G.4

Recall (39) that

$$\underline{\nu}^{\mathrm{ref}}-\frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h},\underline{V}_{h+1}^{\mathrm{ref},\ell_i})=-\frac{1}{n_h^k}(\underline{\chi}_6+\underline{\chi}_7+\underline{\chi}_8),$$

where

$$\begin{aligned}
\underline{\chi}_6&=\sum_{i=1}^{n_h^k}\left((P_{s_h^k,a_h^k,b_h^k,h}(\underline{V}_{h+1}^{\mathrm{ref},\ell_i})^2-(\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}))^2\right),\\
\underline{\chi}_7&=\frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i})\right)^2-\frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2,\\
\underline{\chi}_8&=\frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2-\sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i})^2.
\end{aligned}$$

By Azuma's inequality, with probability at least $1-2\delta$, it holds that

$$|\underline{\chi}_6|\leq H^2\sqrt{2n_h^k\iota},\quad |\underline{\chi}_7|\leq 2H^2\sqrt{2n_h^k\iota}.$$

The term $\underline{\chi}_8$ is bounded slightly differently from χ_8 as follows:

$$-\underline{\chi}_8=\sum_{i=1}^{n_h^k}\left(P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2-\frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k}P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2$$

$$\leq \sum_{i=1}^{n_h^k} \left(P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{REF}}\right)^2 - \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2 \tag{84}$$

$$= \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{REF}}\right)^2 - \frac{1}{n_h^k}\left(\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\underline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)^2$$

$$\leq 2H^2 \sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\lambda_{h+1}^{\ell_i}$$

$$= 2H^2\left(\sum_{i=1}^{n_h^k}\lambda_{h+1}^{\ell_i}(s_{h+1}^{\ell_i}) + \sum_{i=1}^{n_h^k}(P_{s_h^k,a_h^k,b_h^k,h} - \mathbf{1}_{s_{h+1}^{\ell_i}})\lambda_{h+1}^{\ell_i}\right)$$

$$\leq 2H^2 SN_0 + 2H^2\sqrt{2n_h^k\iota}, \tag{85}$$

where (84) follows from the fact that $\underline{V}_{h+1}^{\mathrm{ref},k} \leq \underline{V}_{h+1}^{\mathrm{REF}}$ for any k, h, and (85) holds with probability at least $1-\delta$ due to Azuma's inequality. Therefore,

$$\underline{\nu}_h^{\mathrm{ref},k} - \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, \underline{V}_{h+1}^{\mathrm{ref},\ell_i}) \leq \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}}. \tag{86}$$

By a similar argument as in Appendix G.3.2, we can obtain the desired result

$$\underline{\nu}_h^{\mathrm{ref},k} - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, V_{h+1}^{\pi^k})$$

$$= \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, \underline{V}_{h+1}^{\mathrm{ref},\ell_i}) - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, V_{h+1}^{\pi^k})\right) + \left(\underline{\nu}_h^{\mathrm{ref},k} - \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, \underline{V}_{h+1}^{\mathrm{ref},\ell_i})\right)$$

$$\leq \frac{1}{n_h^k}\sum_{i=1}^{n_h^k}\left(\mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, \underline{V}_{h+1}^{\mathrm{ref},\ell_i}) - \mathbb{V}(P_{s_h^k,a_h^k,b_h^k,h}, V_{h+1}^{\pi^k})\right) + \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}} \tag{87}$$

$$\leq \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}\left|P_{s_h^k,a_h^k,b_h^k,h}(\underline{V}_{h+1}^{\mathrm{ref},\ell_i} - V_{h+1}^{\pi^k})\right| + \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}}$$

$$= \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}\left|P_{s_h^k,a_h^k,b_h^k,h}(\underline{V}_{h+1}^{\mathrm{ref},\ell_i} - V_{h+1}^{\pi^k} + V_{h+1}^* - V_{h+1}^*) - H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right) + H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)\right|$$
$$+ \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}}$$

$$\leq \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}(V_{h+1}^* - \underline{V}_{h+1}^{\mathrm{ref},\ell_i}) + \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k} P_{s_h^k,a_h^k,b_h^k,h}\left(V_{h+1}^* - V_{h+1}^{\pi^k} + H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)\right)$$
$$+ \frac{4H^2}{n_h^k}\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right) + \frac{2H^2SN_0}{n_h^k} + 8H^2\sqrt{\frac{\iota}{n_h^k}} \tag{88}$$

$$\leq \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}(V_{h+1}^* - \underline{V}_{h+1}^{\mathrm{ref},\ell_i})(s_{h+1}^{\ell_i}) + \frac{4H}{n_h^k}\sum_{i=1}^{n_h^k}\left((V_{h+1}^* - V_{h+1}^{\pi^k})(s_{h+1}^{\ell_i}) + H\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)\right)$$
$$+ \frac{4H^2}{n_h^k}\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right) + \frac{2H^2SN_0}{n_h^k} + 20H^2\sqrt{\frac{\iota}{n_h^k}} \tag{89}$$

$$\leq \left(4H\beta + \frac{4H^2SN_0}{n_h^k}\right) + \frac{8H^2}{n_h^k}\left(\beta + \frac{HSN_0}{\check{n}_h^k}\right)$$

$$+\frac{4H^2}{n_h^k}\left(\beta+\frac{HSN_0}{\check{n}_h^k}\right)+\frac{2H^2SN_0}{n_h^k}+20H^2\sqrt{\frac{\iota}{n_h^k}} \tag{90}$$

$$=4H\beta+\frac{12H^2\beta}{n_h^k}+20H^2\sqrt{\frac{\iota}{n_h^k}}+\frac{6H^2SN_0}{n_h^k}+\frac{12H^3SN_0}{n_h^k\check{n}_h^k}$$

$$\le 4H\beta+\frac{12H^2\beta+18H^3SN_0}{n_h^k}+20H^2\sqrt{\frac{\iota}{n_h^k}},$$

where (87) follows from (86), (88) follows from Lemma D.1 and Lemma G.1, (89) holds with probability at least $1-2\delta$ by Azuma's inequality, and (90) follows from Lemma D.1 and Lemma G.1.

G.5. Summarizing Terms T_1-T_4 Together

Recall that $\beta=\frac{1}{\sqrt{H}}$, and $N_0=\frac{c_4SABH^5\iota}{\beta^2}=O(SABH^6\iota)$. By combining (54), (58), (64), (68) and (83), we conclude that with probability at least $1-O(H^2T^4)\delta$, the following bound holds:

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}(1+\frac{1}{H})^{h-1}\Lambda_{h+1}^k\\
&\le O(\log T)\cdot(H^2SN_0+H\sqrt{T\iota})+O(H\sqrt{SABT\iota})\\
&\quad+O\big(\sqrt{SABH^2T\iota}+\sqrt{SABH^2\beta T\iota}+\sqrt{SABH^3\beta^2T\iota}\\
&\quad+S^{\frac{3}{2}}ABH^3N_0^{\frac{1}{2}}\iota\log T+SABH^2\beta^{\frac{1}{2}}\iota^{\frac{1}{2}}\log T+(SAB\iota)^{\frac{3}{4}}H^{\frac{5}{2}}T^{\frac{1}{4}}\big)\\
&=O\left(\sqrt{SABH^2T\iota}+H\sqrt{T\iota}\log T+(SAB\iota)^{\frac{3}{4}}H^{\frac{5}{2}}T^{\frac{1}{4}}\right)\\
&\quad+O\left(\sqrt{SABH^2\beta T\iota}+\sqrt{SABH^3\beta^2T\iota}+SABH^2\beta^{\frac{1}{2}}\iota^{\frac{1}{2}}\log T\right)\\
&\quad+O\left((H^2SN_0+S^{\frac{3}{2}}ABH^3N_0^{\frac{1}{2}}\iota)\log T\right)\\
&=O\left(\sqrt{SABH^2T\iota}+H\sqrt{T\iota}\log T+S^2(AB)^{\frac{3}{2}}H^8\iota^{\frac{3}{2}}T^{\frac{1}{4}}\right).
\end{aligned} \tag{91}$$

H. Proof of Lemma D.6 (Final Step)

Our construction of the correlated policy is inspired by the "certified policies" in [Bai et al., 2020].

Based on the trajectory of the distributions $\{\pi_h^k\}_{h\in[H],k\in[K]}$ specified by Algorithm 3, we construct a correlated policy $\widehat{\pi}_h^k=\widehat{\mu}_h^k\times\widehat{\nu}_h^k$ for each $(h,k)\in[H]\times[K]$. The max-player's policies $\widehat{\mu}_h^k$ and $\widehat{\mu}_{h+1}^k[s,a,b]$ are defined in Algorithm 4, and the min-player's policies can be defined similarly. Further, we define the final output policy π^{out} in Algorithm 2, which first uniformly samples an index k from $[K]$, and then proceeds with $\widehat{\pi}_1^k$. We remark that based on Algorithm 4 and Algorithm 5, the policies $\widehat{\mu}_h^k,\widehat{\nu}_h^k,\widehat{\mu}_{h+1}^k[s,a,b],\widehat{\nu}_{h+1}^k[s,a,b]$ do not depend on the history before step h. Therefore, the action-value functions are well-defined for the corresponding steps.

Algorithm 4 Certified policy $\widehat{\mu}_h^k$ (max-player version)

1: Initialize $k'\leftarrow k$.
2: **for** step $h'\leftarrow h,h+1,\ldots,H$ **do**
3: Receive $s_{h'}$, and take action $a_{h'}\sim\mu_h^{k'}(\cdot|s_{h'})$.
4: Observe $b_{h'}$, and sample $j\leftarrow\text{Unif}([N_{h'}^{k'}(s_{h'},a_{h'},b_{h'})])$
5: Set $k'\leftarrow\check{\ell}_{h',j}^{k'}$.
6: **end for**

In order to show Lemma D.6, it suffices to show the following inequalities

$$\overline{Q}_h^k(s,a,b)\ge Q_h^{\dagger,\widehat{\nu}_{h+1}^k[s,a,b]}(s,a,b),\quad \overline{V}_h^k(s)\ge V_h^{\dagger,\widehat{\nu}_h^k}(s),$$
$$\underline{Q}_h^k(s,a,b)\ge Q_h^{\widehat{\mu}_{h+1}^k[s,a,b],\dagger}(s,a,b),\quad \underline{V}_h^k(s)\ge V_h^{\widehat{\mu}_h^k,\dagger}(s).$$

Algorithm 5 Policy $\widehat{\mu}_{h+1}^k[s,a,b]$ (max-player version)

Sample $j \leftarrow \mathrm{Unif}([N_h^k(s,a,b)])$
$k' \leftarrow \check{\ell}_{h,j}^k$.
for step $h' \leftarrow h+1,\ldots,H$ **do**
  Receive $s_{h'}$, and take action $a_{h'} \sim \mu_h^{k'}(\cdot|s_{h'})$.
  Observe $b_{h'}$, and sample $j \leftarrow \mathrm{Unif}([N_{h'}^{k'}(s_{h'},a_{h'},b_{h'})])$
  Set $k' \leftarrow \check{\ell}_{h',j}^{k'}$.
end for

due to the definition of output policy in Algorithm 2.

Consider a fixed tuple (s,a,b,h,k). Note that the result clearly holds for any s,a,b that is in its first stage, due to our initialization of $\overline{Q}_h^k(s,a,b), \underline{Q}_h^k(s,a,b)$ and $\overline{V}_h^k(s), \underline{V}_h^k(s)$. In the following, we focus on the case where those values have been updated at least once before the k-th episode.

Our proof is based on induction on k. Note first that the claim clearly holds for $k=1$. For $k \geq 2$, assume the claim holds for all $u \in [1:k-1]$. If those values are not updated in the k-th episode, then the claim clearly holds.In the following, we consider the case where those values has just been updated.

(I) We show $\overline{Q}_h^k(s,a,b) \geq Q_h^{\dagger,\widehat{\nu}_{h+1}^k[s,a,b]}(s,a,b)$.

Recall the update rule of the optimistic action-value function

$$\overline{Q}_h(s,a,b) \leftarrow \min\left\{r_h(s,a,b) + \frac{\check{\overline{v}}}{\check{n}} + \gamma, r_h(s,a,b) + \frac{\overline{\mu}^{\mathrm{ref}}}{n} + \frac{\check{\overline{\mu}}}{\check{n}} + \overline{\beta}, \overline{Q}_h^k(s,a,b)\right\}.$$

Besides the last term, there are two non-trivial cases and we will show both of the first two terms are lower-bounded by $Q_h^{\dagger,\widehat{\nu}_{h+1}^k[s,a,b]}(s,a,b)$.

For the first case, we have

$$\begin{aligned}
\overline{Q}_h^k(s,a,b) &= r_h(s,a,b) + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} \overline{V}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) + \gamma_h^k \\
&\geq r_h(s,a,b) + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} \overline{V}_{h+1}^{\dagger,\widehat{\nu}_{h+1}^{\check{\ell}_i}}(s_{h+1}^{\check{\ell}_i}) + \gamma_h^k && (92)\\
&\geq \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} \overline{Q}_h^{\dagger,\widehat{\nu}_{h+1}^{\check{\ell}_i}}(s,a,b) && (93)\\
&\geq \sup_{\mu} \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k} \overline{Q}_h^{\mu,\widehat{\nu}_{h+1}^{\check{\ell}_i}}(s,a,b) && (94)\\
&\geq \overline{Q}_h^{\dagger,\widehat{\nu}_{h+1}^k[s,a,b]}(s,a,b), && (95)
\end{aligned}$$

where (92) follows from the induction hypothesis, (93) follows from the Azuma's inequality, (94) follows from the fact that taking the maximum out of the summation does not increase the sum, and (95) follows from the construction of policy $\widehat{\nu}_{h+1}^k[s,a,b]$ (obtained via the min-player's counterpart of Algorithm 5).

For the second case,

$$\overline{Q}_h^k(s,a,b) = r_h(s,a,b) + \frac{1}{n_h^k}\sum_{i=1}^{n_h^k} \overline{V}_{h+1}^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) + \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\left(\overline{V}_{h+1}^{\check{\ell}_i} - \overline{V}_{h+1}^{\mathrm{ref},\check{\ell}_i}\right)(s_{h+1}^{\check{\ell}_i}) + \overline{\beta}_h^k$$

$$\geq r_h(s,a,b) + P_h\left(\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\overline{V}_{h+1}^{\check{\ell}_i}\right)(s,a,b) + \chi_1 + \chi_2 + \overline{\beta}_h^k$$

$$\geq r_h(s,a,b) + P_h\left(\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\overline{V}_{h+1}^{\dagger,\widehat{\nu}_{h+1}^{\check{\ell}_i}}\right)(s,a,b) \tag{96}$$

$$= \frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\overline{Q}_h^{\dagger,\widehat{\nu}_{h+1}^{\check{\ell}_i}}(s,a,b)$$

$$\geq \sup_{\mu}\frac{1}{\check{n}_h^k}\sum_{i=1}^{\check{n}_h^k}\overline{Q}_h^{\mu,\widehat{\nu}_{h+1}^{\check{\ell}_i}}(s,a,b) \tag{97}$$

$$\geq \overline{Q}_h^{\dagger,\widehat{\nu}_{h+1}^k[s,a,b]}(s,a,b), \tag{98}$$

where

$$\chi_1(k,h) = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{V}_h^{\mathrm{ref},\ell_i}(s_{h+1}^{\ell_i}) - \left(P_h\overline{V}_{h+1}^{\mathrm{ref},\ell_i}\right)(s,a,b)\right),$$
$$\overline{W}_{h+1}^{\ell} = \overline{V}_{h+1}^{\ell} - \overline{V}_{h+1}^{\mathrm{ref},\ell}$$
$$\chi_2(k,h) = \frac{1}{\check{n}}\sum_{i=1}^{\check{n}}\left(\overline{W}_{h+1}^{\check{\ell}_i}(s_{h+1}^{\check{\ell}_i}) - \left(P_h\overline{W}_{h+1}^{\check{\ell}_i}\right)(s,a,b)\right).$$

Here, (96) follows from the concentration result $\overline{\beta} \geq \chi_1 + \chi_2$ (see (29)), (97) follows from the fact that taking the maximum out of summation does not increase the sum, and (98) follows from the construction of policy $\widehat{\nu}_{h+1}^k[s,a,b]$ (obtained via the min-player's counterpart of Algorithm 5).

(II) We show $\overline{V}_h^{k+1}(s) \geq V_h^{\dagger,\widehat{\nu}_h^k}(s)$.

Note that

$$\overline{V}_h^k(s) = (\mathbb{D}_{\pi_h^k}\overline{Q}_h^k)(s) \geq \sup_{\mu}(\mathbb{D}_{\mu\times\nu_h^k}\overline{Q}_h^k)(s)$$
$$\geq \sup_{\mu}\mathbb{E}_{a\sim\mu,b\sim\nu_h^k}Q_h^{\dagger,\widehat{\nu}_{h+1}^k[s,a,b]}(s,a,b) = V_h^{\dagger,\check{\nu}_h^k}(s),$$

where the first inequality follows from the property of the CCE oracle and the second inequality follows from the induction hypothesis.

The other side of bounds can be proved similarly for $\underline{Q}_h^k(s,a,b)$, $Q_h^{\check{\mu}_{h+1}^k[s,a,b],\dagger}(s,a,b)$, $\underline{V}_h^k(s)$, and $Q_h^{\check{\mu}_{h+1}^k,\dagger}(s)$.

I. Supporting Lemmas

Lemma I.1 (Azuma-Hoefdding's inequality)**.** *Suppose $\{X_k\}_{k\geq 0}$ is a martingale and $|X_k - X_{k-1}| \leq c_k$ almost surely. Then, for all positive integers N and all positive ϵ, it holds that*

$$\mathbb{P}[|X_N - X_0| \geq \epsilon] \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^{N}c_k^2}\right).$$

Lemma I.2 (Lemma 10 in [Zhang et al., 2020b])**.** *Let $\{M_n\}_{n\geq 0}$ be martingale such that $M_0 = 0$ and $|M_n - M_{n-1}| \leq c$ for some $c > 0$ and any $n \geq 1$. Let $\mathrm{Var}_n = \sum_{k=1}^{n}\mathbb{E}[(M_k - M_{k-1})^2|\mathcal{F}_{k-1}]$ for $n \geq 0$, where $\mathcal{F}_k = \sigma(M_1, M_2, \ldots, M_k)$. Then for any positive integer n, and any $\epsilon, p > 0$, we have*

$$\mathbb{P}\left[|M_n| \geq 2\sqrt{\mathrm{Var}_n\log\frac{1}{p}} + 2\sqrt{\epsilon\log\frac{1}{p}} + 2c\log\frac{1}{p}\right] \leq \left(\frac{2nc^2}{\epsilon} + 2\right)p.$$

Lemma I.3 (Variant of Lemma 11 in [Zhang et al., 2020b]). *For any $\alpha \in (0,1)$ and non-negative weights $\{w_h(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A},b\in\mathcal{B},h\in[H]}$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k,b_h^k)}{(n_h^k)^\alpha} \le \frac{2^\alpha}{1-\alpha}\sum_{s,a,b,h} w_h(s,a,b)(N_h^{K+1}(s,a,b))^{1-\alpha},$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k,b_h^k)}{(\check{n}_h^k)^\alpha} \le \frac{2^{2\alpha}H^\alpha}{1-\alpha}\sum_{s,a,b,h} w_h(s,a,b)(N_h^{K+1}(s,a,b))^{1-\alpha}.$$

In the case $\alpha = 1$, it holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k,b_h^k)}{n_h^k} \le 2\sum_{s,a,b,h} w_h(s,a,b)\log(N_h^{K+1}(s,a,b)),$$
$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{w_h(s_h^k,a_h^k,b_h^k)}{\check{n}_h^k} \le 4H\sum_{s,a,b,h} w_h(s,a,b)\log(N_h^{K+1}(s,a,b)).$$