Sample-Efficient Tabular Self-Play for Offline Robust Reinforcement Learning

Na Li
Zhejiang University
nlee@zju.edu.cn

Zewu Zheng
The Chinese University of Hong Kong
zhengzw@link.cuhk.edu.hk

Wei Ni
Edith Cowan University and University of New South Wales
wei.ni@ieee.org

Hangguan Shan
Zhejiang University
hshan@zju.edu.cn

Wenjie Zhang
University of New South Wales
wenjie.zhang@unsw.edu.au

Xinyu Li
Huazhong University of Science and Technology
lixinyu@hust.edu.cn

Abstract

Multi-agent reinforcement learning (MARL), as a thriving field, explores how multiple agents independently make decisions in a shared dynamic environment. Due to environmental uncertainties, policies in MARL must remain robust to tackle the sim-to-real gap. We focus on robust two-player zero-sum Markov games (TZMGs) in offline settings, specifically on tabular robust TZMGs (RTZMGs). We propose a model-based algorithm (*RTZ-VI-LCB*) for offline RTZMGs, which is optimistic robust value iteration combined with a data-driven Bernstein-style penalty term for robust value estimation. By accounting for distribution shifts in the historical dataset, the proposed algorithm establishes near-optimal sample complexity guarantees under partial coverage and environmental uncertainty. An information-theoretic lower bound is developed to confirm the tightness of our algorithm's sample complexity, which is optimal regarding both state and action spaces. To the best of our knowledge, RTZ-VI-LCB is the first to attain this optimality, sets a new benchmark for offline RTZMGs, and is validated experimentally.

1 Introduction

Multi-agent reinforcement learning (MARL), which focuses on developing algorithms that enable multiple agents to learn and make decisions in dynamic environments, has garnered significant attention in gaming [35] and autonomous driving [4]. Offline MARL, addresses the high cost of interacting with the environment by leveraging historical data collected from past interactions generated under unknown or biased behavior policies [22]. The dynamic and non-stationary nature of real-world environments introduces critical uncertainties. Robustness becomes important in ensuring stable decision-making, because standard MARL algorithms under ideal conditions are highly sensitive and prone to catastrophic failures when faced with even minor adversarial perturbations [48, 45, 46]. In this sense, robust guarantees are particularly vital in offline MARL, highlighting the core of *offline robust MARL*. Two-player zero-sum Markov games (TZMGs) represent a compelling setting of MARL, giving rise to the field of robust TZMGs (RTZMGs) from robust MARL.

A key challenge in offline RTZMGs is addressing environmental uncertainties with as few samples as possible under partial and limited coverage. Historical data often only offers partial and limited coverage of the state-action space, leading to poor estimates of model parameters and unreliable policy. Besides, environmental uncertainties arise from model mismatches, system noise, and the disparity between simulation and real-world scenarios. To address uncertainties, RTZMGs incorporate

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Table 1: A comparison between RTZ-VI-LCB and $\mathrm{P}^2\mathrm{M}^2\mathrm{PO}$ [5] on finding an ε-optimal robust Nash policy in finite-horizon offline RTZMGs with $f(\sigma^+,\sigma^-,H) = \min\left\{\frac{(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2}, \frac{(H\sigma^--1+(1-\sigma^-)^H)}{(\sigma^-)^2}, H\right\}$, where the uncertainty set is quantified by total variation (TV) distance. The sample complexities omit all logarithmic factors.

Algorithm	Sample complexity	Uncertainty level
$\mathrm{P}^2\mathrm{M}^2\mathrm{PO}$ [5]	$\frac{C_{\mathrm{r}}H^5S^2AB}{\varepsilon^2}$	not considered
RTZ-VI-LCB (Ours)	$\frac{C_{\mathrm{r}}^\star H^4S(A+B)}{\varepsilon^2}f(\sigma^+,\sigma^-,H)$	full range
Lower bound	$\frac{C_{\mathrm{r}}^\star SH^4(A+B)}{\varepsilon^2}$	$\min\{\sigma^+,\sigma^-\} \lesssim \frac{1}{H}$
Lower bound	$\frac{C_{\mathrm{r}}^\star SH^3(A+B)}{\varepsilon^2\min\{\sigma^+,\sigma^-\}}$	$\min\{\sigma^+,\sigma^-\} \gtrsim \frac{1}{H}$

equilibria not only between the two players but also with their adversarial strategies, considering worst-case environments selected from predefined uncertainty sets for each player.

Despite recent efforts [21, 5, 48, 29], there remains a fundamental gap in learning effectively in offline RTZMGs, primarily due to high sample complexity. For a tabular RTZMG (formal definition in Section 2) with horizon length H, states S, actions $\{A, B\}$, and uncertainty levels $\{\sigma^+, \sigma^-\}$ for the two players, the best sample complexity for ε-optimal robust Nash equilibrium (NE) in the offline setting to date is $\widetilde{O}\left(\frac{C_{\mathrm{r}}H^5S^2AB}{\varepsilon^2}\right)$ achieved by $\mathrm{P}^2\mathrm{M}^2\mathrm{PO}$ [5], which demonstrates near-optimal sample complexity in H, S, and A, B, but overlooks the influence of uncertainty levels and faces the curse of multiagency [40]. Hence, the key research question addressed in this paper is

Can we design an efficient algorithm for offline RTZMGs with partial state-action coverage while ensuring robustness to uncertainties?

1.1 Contribution

In this paper, we design a novel model-based algorithm *RTZ-VI-LCB* for offline RTZMGs, which is an optimistic variant of robust value iteration. RTZ-VI-LCB involves a data-informed Bernstein-style penalty for robust value estimation to effectively capture the variance structure, and a two-stage subsampling method to suppress the statistical dependencies of the historical data. Notably, this is the *first time* that the optimal dependence of sample complexity on S and $\{A, B\}$ and the best dependence on H has been achieved for offline RTZMGs. Table 1 compares the sample complexity between our approach and the status quo. The main contributions are outlined as follows.

(i) Robust unilateral clipped concentrability: With this new criterion, we define a measure $C_{\mathrm{r}}^\star \in \left[\frac{1}{S(A+B)}, \infty\right)$ for the quality of historical data. It captures the distribution shift between the behavior policy $(\mu^{\mathsf{n}}, \nu^{\mathsf{n}})$ and the single optimal robust policies $(\mu, \nu^\star)$ and $(\mu^\star, \nu)$ under environmental uncertainty in the partial coverage, and offers a tighter measure of distribution mismatch than C_{r} used in $\mathrm{P}^2\mathrm{M}^2\mathrm{PO}$ [5]. The introduction of $C_{\mathrm{r}}^\star$ improves sample complexity.

(ii) Near-optimal sample complexity upper bound: RTZ-VI-LCB can provably find an ε-optimal robust NE policy as long as the sample size exceeds $\widetilde{O}\left(\frac{C_{\mathrm{r}}^\star H^4S(A+B)}{\varepsilon^2}f(\sigma^+,\sigma^-,H)\right)$ with an ε-independent burn-in cost. This significantly improves upon the prior art [5] on state S and action $\{A, B\}$, and further delineates the impact of the uncertainty levels $\{\sigma^+, \sigma^-\}$.

(iii) Information-theoretic sample complexity lower bound: We establish a tight lower bound for RTZMGs, revealing at least $\Omega\left(C_{\mathrm{r}}^\star SH^4(A+B)/\varepsilon^2\right)$ samples are needed for an ε-optimal robust NE policy under the uncertainty level $\min\{\sigma^+,\sigma^-\} \lesssim \frac{1}{H}$, and at least $\Omega\left(C_{\mathrm{r}}^\star SH^3(A+B)/(\varepsilon^2\min\{\sigma^+,\sigma^-\})\right)$ samples are needed under $\min\{\sigma^+,\sigma^-\} \gtrsim \frac{1}{H}$. Comparing these upper and lower bounds confirms the optimality of RTZ-VI-LCB w.r.t. state S and actions $\{A, B\}$ in sample complexity across uncertainty levels.

(iv) Extension to multi-agent RL: We generalize RTZ-VI-LCB to *Multi-RTZ-VI-LCB* for robust multi-player general-sum Markov games, and achieve an ε-optimal robust NE policy for $\widetilde{O}\left(C_{\mathrm{r}}^\star H^4S\sum_{i=1}^m A_i \min\left\{\left\{(H\sigma_i-1+(1-\sigma_i)^H)/(\sigma_i)^2\right\}_{i=1}^m, H\right\}/\varepsilon^2\right)$ samples with M players, A_i actions, and uncertainty level σ_i per player.

1.2 Related Work

This section reviews a curated selection of related research focusing on provably tabular RL.

Finite-sample studies of standard TZMGs. Markov games (MGs), or stochastic games, were proposed in the early 1950s [32]. Then, extensive research has been conducted, and MARL has gained significant attention [30], particularly around Nash equilibrium [27, 23]. Numerous MARL algorithms with provable convergence and asymptotic guarantees have been developed [31]. More recent work has focused on creating algorithms for standard MARL with non-asymptotic guarantees through finite-sample analysis. In this area, most efforts to compute Nash equilibria are focused on TZMGs. The studies in [2] and [41] were the first to provide non-asymptotic sample complexity guarantees for model-based (e.g., VI-Explore and VI-ULCB) and model-free algorithms (e.g., OMNI-VI). Further improvements in sample complexity have been explored [10, 8, 28, 13, 26].

Robustness in MARL. Although progress has been made in MARL, existing algorithms may struggle when faced with environmental uncertainties, leading to significantly deviated equilibria. MARL robustness against uncertainties has drawn attention in different parts of MGs [38], including state [49], environment (reward and transition dynamics), agent types [47], and other agents' policies [20]. A typical method to address robustness against uncertainties of the environment is distributionally robust optimization (DRO), which is a method predominantly explored in supervised learning [3, 14, 6]. The application of DRO to manage model uncertainty in single-agent RL [17] has attracted considerable attention. However, when extended to MARL, researchers formulated the problem as robust MGs armed with DRO and developed a relatively understudied field with only a few proved algorithms [5, 21, 29, 48, 34]. Thus, relevant algorithms based on partial coverage of datasets while considering the uncertainty level are lacking.

Single-agent robust offline RL. In single-agent offline RL, addressing uncertainties of environments using DRO—such as robust Markov decision processes (MDPs) and distributionally robust dynamic programming—has attracted considerable interest in both theoretical research and practical applications [16]. Recent work has focused on the finite-sample performance of provable robust offline RL algorithms, exploring different divergence functions for uncertainty sets, various sampling mechanisms, and related challenges [5]. It has been shown that addressing robust MDPs does not demand more samples compared with those needed for standard MDPs [33]. However, RTZMGs present additional complexities beyond those in robust single-agent offline RL.

2 Problem Formulation

This paper focuses on offline RTZMGs, which is a robust version of standard offline TZMGs by taking environmental uncertainties into consideration. RTZMGs form a broader class than standard TZMGs, accommodating various prescribed environmental uncertainty sets. In RTZMGs, the dynamics extend standard TZMGs by incorporating two players, as well as their respective situations that determine the worst-case transitions. We investigate an efficient algorithm to achieve robustness and optimal sample complexity on state S and actions $\{A, B\}$ under partial coverage of the state-action space.

An RTZMG under the finite-horizon setting can be defined as $\mathcal{MG}_{\mathsf{r}} = \left\{\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{U}_{\rho}^{\sigma^+}\left(P^0\right), \mathcal{U}_{\rho}^{\sigma^-}\left(P^0\right), r, H\right\}$, where $\mathcal{S} \coloneqq \{1, \cdots, S\}$ is the state space of size S; $(\mathcal{A} \coloneqq \{1, \cdots, A\}, \mathcal{B} \coloneqq \{1, \cdots, B\})$ denotes the action spaces of the max-player and the min-player with sizes A and B, respectively; H is the horizon length; $r = \{r_h\}_{h=1}^{H}$ represents the immediate reward obtained at time step h. Specifically, $r_h(s, a, b)$ is assumed to be deterministic on a state-action pair (s, a, b) and falls within the range $[0, 1]$. This reward can represent both the gain of the max-player and the loss of the min-player. Here, $\mathcal{U}_{\rho}^{\sigma^+}(P^0)$ and $\mathcal{U}_{\rho}^{\sigma^-}(P^0)$ represent the uncertainty sets for the max-player and min-player, respectively.

Unlike standard TZMGs that assume a fixed transition kernel, these uncertainty sets account for bounded perturbations in the transition kernel and enable modeling of environmental uncertainties. These uncertainty sets are centered on a nominal kernel $P^0 : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto \Delta(\mathcal{S})$, with their size and shape defined by a distance metric ρ and radius parameters $\sigma^+ > 0$ and $\sigma^- > 0$. Considering players' individual properties, both players can independently define their uncertainty sets $\mathcal{U}_{\rho}^{\sigma^+}(P^0)$

and $\mathcal{U}_\rho^{\sigma^-}(P^0)$, by specifying different sizes ($\sigma^+ > 0$ and $\sigma^- > 0$) and potentially using distinct divergence functions (ρ) to shape these sets. For illustration convenience, we consider the same divergence function for both players in this paper.

Uncertainty set with *two-player* (s,a,b)-*rectangularity*. According to the transition kernel uncertainty sets $\mathcal{U}_\rho^{\sigma^+}(P^0)$ and $\mathcal{U}_\rho^{\sigma^-}(P^0)$ defined above, we adapt the *rectangularity* condition to a two-player setting inspired by [33, 17], termed *two-player-wise* (s,a,b)-*rectangularity*. The adaptation enhances computational tractability and facilitates the robust version of Bellman recursions. It permits each player to select its uncertainty set independently, which can be decomposed for each state-action pair into a product of subsets. Thus, the uncertainty sets $\mathcal{U}_\rho^{\sigma^+}(P^0)$ and $\mathcal{U}_\rho^{\sigma^-}(P^0)$ for the two players, adhering to *two-player-wise* (s,a,b)-*rectangularity*, are mathematically defined as

$$\mathcal{U}_\rho^{\sigma^+}\left(P^0\right) := \otimes\, \mathcal{U}_\rho^{\sigma^+}\left(P^0_{h,s,a,b}\right), \quad \mathcal{U}_\rho^{\sigma^-}\left(P^0\right) := \otimes\, \mathcal{U}_\rho^{\sigma^-}\left(P^0_{h,s,a,b}\right), \tag{1}$$

where $\mathcal{U}_\rho^{\sigma^+}\left(P^0_{h,s,a,b}\right) := \left\{P_{h,s,a,b} \in \Delta(\mathcal{S}) : \rho\left(P_{h,s,a,b}, P^0_{h,s,a,b}\right) \le \sigma^+\right\}$, $\otimes$ represents the Cartesian product, and $\mathcal{U}_\rho^{\sigma^-}\left(P^0_{h,s,a,b}\right)$ can be defined similarly. We define a vector of the transition kernel P or P^0 at any state-action pair (s,a,b) as

$$P_{h,s,a,b} := P_h(\cdot \,|\, s,a,b) \in \mathbb{R}^{1\times S}, \quad P^0_{h,s,a,b} := P^0_h(\cdot \,|\, s,a,b) \in \mathbb{R}^{1\times S}. \tag{2}$$

Here, the distance function ρ for each player's uncertainty set can be selected from various options that quantify differences between probability vectors. These include f-divergences (e.g., KL divergence, TV distance, and chi-square) [44], the Wasserstein distance [42], and ℓ_q norms [9].

Offline dataset. Let $\mathcal{D}$ be a dataset consisting of K episodes under independence, with each episode produced by implementing a behavior policy $\{\mu_h^{\mathsf{n}}, \nu_h^{\mathsf{n}}\}_{h=1}^H$ in a nominal MDP $\mathcal{M}^0 = \big(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, P^0 := \{P_h^0\}_{h=1}^H, \{r_h\}_{h=1}^H\big)$. For $1 \le k \le K$, the k-th episode $\big(s_1^k, a_1^k, b_1^k, \ldots, s_H^k, a_H^k, b_H^k, s_{H+1}^k\big)$ is generated as follows:

$$s_1^k \sim \varrho^{\mathsf{n}}, \quad a_h^k \sim \mu_h^{\mathsf{n}}(\cdot \,|\, s_h^k), \quad b_h^k \sim \nu_h^{\mathsf{n}}(\cdot \,|\, s_h^k), \quad s_{h+1}^k \sim P_h^0(\cdot \,|\, s_h^k, a_h^k, b_h^k), \quad 1 \le h \le H. \tag{3}$$

Throughout this paper, ϱ^{n} denotes the initial distribution related to a historical dataset. We use the short-hand notation for the occupancy distribution with respect to (w.r.t.) the behavior policy $(\mu^{\mathsf{n}}, \nu^{\mathsf{n}})$ as: $\forall (h,s,a,b) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B}$,

$$d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}},P^0}(s) := \mathbb{P}(s_h = s | s_1 \sim \varrho^{\mathsf{n}}, \mu^{\mathsf{n}}, \nu^{\mathsf{n}}, P^0);\ d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}},P^0}(s,a,b) := d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}},P^0}(s)\mu_h^{\mathsf{n}}(a \,|\, s)\, \nu_h^{\mathsf{n}}(b \,|\, s), \tag{4}$$

which are simplified to $d_h^{\mathsf{n},P^0}(s) = d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}},P^0}(s)$ and $d_h^{\mathsf{n},P^0}(s,a,b) = d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}},P^0}(s,a,b)$. Similarly, for any product policy (μ,ν), we define: $\forall (h,s,a,b) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B}$

$$d_h^{\mu,\nu,P}(s) := \mathbb{P}(s_h = s \,|\, s_1 \sim \varrho, \mu, \nu, P);\ d_h^{\mu,\nu,P}(s,a,b) := d_h^{\mu,\nu,P}(s)\mu_h(a \,|\, s)\, \nu_h(b \,|\, s). \tag{5}$$

Robust value functions. In RTZMGs, players seek to optimize their worst-case performance across all possible transition kernels within their respective uncertainty sets $\mathcal{U}_\rho^{\sigma^+}\left(P^0\right)$ and $\mathcal{U}_\rho^{\sigma^-}\left(P^0\right)$. For any product policy $(\mu \times \nu) \in \Delta(\mathcal{A} \times \mathcal{B})$, the max-player's worst-case performance at time step h is measured with the *robust value function* V_h^{μ,ν,σ^+} and the *robust Q-function* Q_h^{μ,ν,σ^+}, $\forall (h,s,a,b) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, as given by

$$V_h^{\mu,\nu,\sigma^+}(s) := \inf\nolimits_{P \in \mathcal{U}_\rho^{\sigma^+}(P^0)} V_h^{\mu,\nu,P}(s), \quad Q_h^{\mu,\nu,\sigma^+}(s,a,b) := \inf\nolimits_{P \in \mathcal{U}_\rho^{\sigma^+}(P^0)} Q_h^{\mu,\nu,P}, \tag{6a}$$

$$V_h^{\mu,\nu,\sigma^-}(s) := \sup\nolimits_{P \in \mathcal{U}_\rho^{\sigma^-}(P^0)} V_h^{\mu,\nu,P}(s), \quad Q_h^{\mu,\nu,\sigma^-}(s,a,b) := \sup\nolimits_{P \in \mathcal{U}_\rho^{\sigma^-}(P^0)} Q_h^{\mu,\nu,P}, \tag{6b}$$

where

$$V_h^{\mu,\nu,P}(s) := \mathop{\mathbb{E}}_{\mu,\nu,P}\left[\sum\nolimits_{t=h}^H r_t\big(s_t,a_t,b_t\big) \mid s_h = s\right];$$

$$Q_h^{\mu,\nu,P}(s,a,b) := \mathop{\mathbb{E}}_{\mu,\nu,P}\left[\sum\nolimits_{t=h}^H r_t\big(s_t,a_t,b_t\big) \mid s_h = s, a_h = a, b_h = b\right].$$

Robust Bellman equations. Based on the robust value functions in (6), RTZMGs include a robust version of the Bellman equation, dubbed *robust Bellman equation*. The robust value functions $V_h^{\mu,\nu,\sigma^+}(s)$ for the max-player with any product policy (μ,ν) satisfy: $\forall(h,s)\in[H]\times\mathcal{S}$,

$$V_h^{\mu,\nu,\sigma^+}(s) = \mathbb{E}_{(a,b)\sim(\mu_h(a),\nu_h(a))}\Big[r_h(s,a,b) + \inf_{P\in\mathcal{U}_\rho^{\sigma^+}(P^0_{h,s,a,b})} PV_{h+1}^{\mu,\nu,\sigma^+}\Big]. \tag{7}$$

Likewise, $V_h^{\mu,\nu,\sigma^-}(s)$ can be obtained for the min-player. Note that the robust Bellman equations are intrinsically connected to the *two-player-wise* (s,a,b)*-rectangularity* condition (see (1)) applied to the uncertainty set. This condition separates the dependencies of uncertainty subsets among different time steps, the players, and state-action pairs, thus leading to the Bellman recursion.

Optimal robust policy. Again, based on (6), we define the maximum robust value function of the max-player under the fixed opponent policy as: $\forall(h,s)\in[H]\times\mathcal{S}$,

$$V_h^{\star,\nu,\sigma^+}(s) \coloneqq \max_{\mu:\mathcal{S}\times[H]\mapsto\Delta(\mathcal{A})} V_h^{\mu,\nu,\sigma^+}(s) = \max_{\mu:\mathcal{S}\times[H]\mapsto\Delta(\mathcal{A})} \inf_{P\in\mathcal{U}_\rho^{\sigma^+}(P^0)} V_h^{\mu,\nu,P}(s).$$

The maximum robust value function for the min-player can be obtained similarly.

As proved in [5], there is at least one policy, denoted by $\mu_h^\star(s):\mathcal{S}\times[H]\mapsto\Delta(\mathcal{A})$ (for the max-player) and $\nu_h^\star(s):\mathcal{S}\times[H]\mapsto\Delta(\mathcal{B})$ (for the min-player), corresponding to the *robust best-response policy*. These policies can simultaneously achieve $V_h^{\star,\nu,\sigma^+}(s)$ (for the max-player) and $V_h^{\mu,\star,\sigma^-}(s)$ (for the min-player) for all $s\in\mathcal{S}$ and $h\in[H]$.

Robust Nash equilibrium. We introduce the robust variant of standard solution concepts—robust NE for RTZMGs. A product policy (μ,ν) is considered a *robust NE* if: $\forall(s)\in\mathcal{S}$

$$V_h^{\star,\nu,\sigma^+}(s) = V_h^{\star,\sigma^+}(s); \quad V_h^{\mu,\star,\sigma^-}(s) = V_h^{\star,\sigma^-}(s). \tag{8}$$

A robust NE signifies that given the product policy (μ,ν) of the opponents, no player can enhance their outcome by deviating from their current policy unilaterally when each player accounts for the worst-case scenario within their uncertainty set $\mathcal{U}_\rho^{\sigma^+}(P^0)$ or $\mathcal{U}_\rho^{\sigma^-}(P^0)$.

Since finding exact robust equilibria can be complex and may not always be feasible, practitioners often seek approximate equilibria. In this context, a product policy $(\mu\times\nu)\in\Delta(\mathcal{A}\times\mathcal{B})$ can be termed an ε*-robust NE* if

$$\mathrm{Gap}(\mu,\nu) \coloneqq \max\{V_1^{\star,\nu,\sigma^+}(\varrho) - V_1^{\star,\sigma^+}(\varrho), \quad V_1^{\star,\sigma^-}(\varrho) - V_1^{\mu,\star,\sigma^-}(\varrho)\} \le \varepsilon, \tag{9}$$

where

$$V_1^{\star,\nu,\sigma^+}(\varrho) = \mathbb{E}_{s\sim\varrho}V_1^{\star,\nu,\sigma^+}(s), \quad V_1^{\star,\sigma^+}(\varrho) = \mathbb{E}_{s\sim\varrho}V_1^{\star,\sigma^+}(s).$$

The definitions of $V_1^{\mu,\star,\sigma^-}(\varrho)$ and $V_1^{\star,\sigma^-}(\varrho)$ can be obtained similarly. The existence of a robust NE has been proved for general divergence functions in the uncertainty set in [5].

Our Goal With a dataset collected from the nominal environment, our objective is to find a solution yielding an ε-robust NE for RTZMG w.r.t. a specified uncertainty set $\mathcal{U}(P^0)$ around the nominal kernel, minimizing the number of samples required under partial coverage of the state-action space.

3 Algorithm Design

In this section, we propose an efficient model-based algorithm, RTZ-VI-LCB, to learn a robust NE policy. Notably, RTZ-VI-LCB achieves a near-optimal sample complexity with robustness, which is designed for offline RTZMGs within the finite-horizon setting.

3.1 Building an Empirical Nominal MDP

According to the empirical frequencies of state transitions, we can construct an empirical estimate $\widehat{P}^0 = \{\widehat{P}_h^0\}_{h=1}^H$ of P^0, where $\forall(h,s,a,b,s')\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times\mathcal{S}$

$$\widehat{P}_h^0(s'\mid s,a,b) = \begin{cases} \frac{\sum_{i=1}^N \mathbb{1}\{(s_i,a_i,b_i,s_i')=(s,a,b,s')\}}{N_h(s,a,b)}, & \text{if } N_h(s,a,b)>0; \\ 1/S, & \text{if } N_h(s,a,b)=0, \end{cases} \tag{10}$$

$$\widehat{r}_h\left(s,a,b\right)=\begin{cases}r_h\left(s,a,b\right), & \text{if } N_h\left(s,a,b\right)>0;\\ 0, & \text{if } N_h\left(s,a,b\right)=0,\end{cases} \tag{11}$$

where $N_h(s,a,b)$ represents the total number of sample transitions from (s,a,b) at step h, and

$$N_h(s,a,b)\coloneqq\sum\nolimits_{i=1}^{N}\mathbb{1}\big\{(s_i,a_i,b_i)=(s,a,b)\big\}. \tag{12}$$

Although it is feasible to decompose the historical dataset $\mathcal{D}$ into sample transitions, the dependencies between transitions within the same episode introduce significant complexities to the analysis.

Algorithm 1 Two-stage subsampling for RTZ-VI-LCB.

input Dataset $\mathcal{D}$, probability δ.

1: **Step 1: Data Partitioning.** Split $\mathcal{D}$ into two equal-sized subsets, $\mathcal{D}^{\mathsf{m}}$ and $\mathcal{D}^{\mathsf{a}}$, each containing $K/2$ trajectories.

2: **Step 2: Defining Transition Bounds.** For step h and state s, denote the number of transitions from $\mathcal{D}^{\mathsf{m}}$ (resp. $\mathcal{D}^{\mathsf{a}}$) as $N_h^{\mathsf{m}}(s)$ (resp. $N_h^{\mathsf{a}}(s)$). Construct the trimmed count as:

$$N_h^{\mathsf{t}}(s)\coloneqq\max\left\{N_h^{\mathsf{a}}(s)-10\sqrt{N_h^{\mathsf{a}}(s)\log\frac{HS}{\delta}},0\right\}; \tag{13}$$

3: **Step 3: Generating Subsampled Dataset.** Randomly sample transitions (quadruples of the form (s,a,b,h,s')) from $\mathcal{D}^{\mathsf{m}}$ uniformly. For each $(s,h)\in\mathcal{S}\times[H]$, include $\min\{N_h^{\mathsf{t}}(s),N_h^{\mathsf{m}}(s)\}$ transitions in the new dataset $\mathcal{D}^{\mathsf{t}}$.

output Set $\mathcal{D}_0=\mathcal{D}^{\mathsf{t}}$.

Inspired by [24], we design a new two-stage subsampling method for RTZMGs, as shown in Algorithm 1. This method effectively reduces statistical dependencies, resulting in a distributionally equivalent dataset $\mathcal{D}_0$ composed of independent samples. The property of $\mathcal{D}_0$ is summarized in the following lemma and formally proved in Appendix C.

Lemma 3.1. *The dataset produced by the two-stage subsampling method is distributionally identical to $\mathcal{D}_0$ with probability of at least $1-8\delta$, where $\{N_h(s,a,b)\}$ are independent of the sample transitions in $\mathcal{D}^0$ and obey: $\forall(h,s,a,b)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}$,*

$$N_h(s,a,b)\geq\frac{Kd_h^{\mathsf{m}}(s,a,b)}{8}-5\sqrt{Kd_h^{\mathsf{m}}(s,a,b)\log\frac{KH}{\delta}}. \tag{14}$$

By applying the two-fold sampling method, we can treat the dataset $\mathcal{D}_0$ as having independent samples, simplifying the analysis significantly as supported by Lemma 3.1.

3.2 Optimistic Variant of Robust Value Iteration with Lower Confidence Bounds

We propose a model-based algorithm, RTZ-VI-LCB, for solving RTZMGs using an approximate $\widehat{P}^0$ for P^0, which is the nominal transition kernel. Specifically, we introduce value iteration with lower confidence bounds for RTZMGs to compute a robust NE for two players, along with a data-informed penalty, as summarized in Algorithm 2.

Our algorithm begins at the final time step $h=H$ and proceeds backward through $h=H-1,H-2,\ldots,1$. Following from single-agent offline RL algorithms [24, 19], we design an optimistic robust Q-value for all $(h,s,a,b)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ as

$$\widehat{Q}_h^{+}\left(s,a,b\right)=\min\left\{\widehat{r}_h\left(s,a,b\right)+\inf\nolimits_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{h,s,a,b}\right)}P\widehat{V}_{h+1}^{+}+\beta_h\left(s,a,b,\widehat{V}_{h+1}^{+}\right),H\right\}; \tag{15a}$$

$$\widehat{Q}_h^{-}\left(s,a,b\right)=\max\left\{\widehat{r}_h\left(s,a,b\right)+\sup\nolimits_{P\in\mathcal{U}^{\sigma^-}\left(\widehat{P}^0_{h,s,a,b}\right)}P\widehat{V}_{h+1}^{-}-\beta_h\left(s,a,b,\widehat{V}_{h+1}^{-}\right),0\right\}, \tag{15b}$$

to estimate the robust Q-function at time step $h\in[H]$ as $\widehat{Q}_h^{+}$ and $\widehat{Q}_h^{-}$.

Dual problem. Solving (15) directly is computationally intensive because it requires optimizing over an S-dimensional probability simplex, which becomes exponentially more difficult as the state

space size S increases. Fortunately, strong duality for TV distance allows us to tackle this problem by solving its dual [17] as

$$\inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{h,s,a,b}\right)} P\widehat{V}^+_{h+1} = \max_{\alpha\in[\min_s \widehat{V}^+_{h+1}, \max_s \widehat{V}^+_{h+1}]} \left\{\widehat{P}^0_{h,s,a,b}\left[\widehat{V}^+_{h+1}\right]_\alpha - \sigma^+\left(\alpha - \min_{s'}\left[\widehat{V}^+_{h+1}\right]_\alpha(s')\right)\right\}, \quad (16)$$

where $\left[\widehat{V}^+_{h+1}\right]_\alpha$ denotes the clipped versions of $\widehat{V}^+_{h+1}\in\mathbb{R}^S$ based on some level $\alpha\geq 0$, as follows.

$$\left[\widehat{V}^+_{h+1}\right]_\alpha(s) := \begin{cases}\alpha, & \text{if } \widehat{V}^+_{h+1}(s)>\alpha;\\ \widehat{V}^+_{h+1}(s), & \text{otherwise.}\end{cases} \quad (17)$$

Moreover, $\sup_{P\in\mathcal{U}^{\sigma^-}\left(\widehat{P}^0_{h,s,a,b}\right)} P\widehat{V}^-_{h+1}$ can be defined similarly. See Appendix B for details.

Penalty term. We design a data-driven penalty term, $\beta_h(s,a,b,\widehat{V})$, to account for uncertainty in value estimates, resulting in an optimistic robust Q-function estimate. To achieve this, we utilize a Bernstein-style penalty, which effectively captures the variance structure over time [24]. In particular, for any $(s,a,b,h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$ and $\delta\in(0,1)$, the penalty term $\beta_h(s,a,b,\widehat{V})$ is defined as

$$\beta_h\left(s,a,b,\widehat{V}\right) = \min\left\{\max\left\{\sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_h(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(\widehat{V})}, \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_h(s,a,b)}\right\}, H\right\}, \quad (18)$$

where C_{n} is some universal constant, and

$$\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}\left(\widehat{V}\right) := \widehat{P}^0_{h,s,a,b}\widehat{V}^2 - (\widehat{P}^0_{h,s,a,b}\widehat{V})^2. \quad (19)$$

Note that we choose $\widehat{P}^0$ in the variance term $\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(\widehat{V})$, as opposed to P^0, since we have no access to the true transition kernel P^0. The penalty term $\beta_h\left(s,a,b,\widehat{V}\right)$ is crafted to address the unique structure of RTZMGs, distinguishing it from the penalty terms used in standard offline TZMGs [10, 24]. In particular, it provides a tight upper bound on statistical uncertainty, considers the non-linear and implicit dependency introduced by the uncertainty set $\mathcal{U}(P^0)$, and addresses challenges not present in standard MDPs.

Policy estimation. We update the policies using the estimated Q-functions with uncertainty, as described in line 5 of Algorithm 2. For any matrix $\mathbf{N}\in\mathbb{R}^{A\times B}$, the function $\mathsf{ComputNash}(\mathbf{N})$ returns a solution $(\widehat{w},\widehat{z})$ to the minimax problem $\max_{w\in\Delta(\mathcal{A})}\min_{z\in\Delta(\mathcal{B})} w^\top\mathbf{N}z$ [11]. In other words, for each $s\in\mathcal{S}$, we compute the NE policies $\left(\mu^+_h(s),\nu^+_h(s)\right)$ and $\left(\mu^-_h(s),\nu^-_h(s)\right)\in\Delta(\mathcal{A})\times\Delta(\mathcal{B})$ for the robust zero-sum matrix games with payoff matrices $\widehat{Q}^+_h(s,\cdot,\cdot)$ and $\widehat{Q}^-_h(s,\cdot,\cdot)$, respectively. Solving robust zero-sum matrix games is generally PPAD-hard as the players can potentially choose different worst-case transition kernels.

4 Performance Guarantees

We provide the theoretical guarantee of RTZ-VI-LCB, including the upper and lower bounds of sample complexity. **We also validate the theoretical guarantee through numerical experiments; see Appendix A.**

Robust unilateral clipped concentrability. For offline RTZMGs, it is essential to measure the distributional discrepancy between the historical data and the target data. Drawing on the *single-policy clipped concentrability* in the single-agent RL [24], we propose a novel criterion, *robust unilateral clipped concentrability*, to measure the distributional discrepancy for RTZMGs:

Algorithm 2 Value iteration with lower confidence bounds for RTZMGs (RTZ-VI-LCB).

1: **Initialization**: Set uncertainty levels σ^- and σ^+; set $\widehat{V}_h^-(s)=0$ and $\widehat{V}_h^+(s)=H$ for all $(s,h)\in\mathcal{S}\times[H+1]$; set $\widehat{Q}_h^-(s,a,b)=0$ and $\widehat{Q}_h^+(s,a,b)=H$ for all $(s,a,b,h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H+1]$.
2: **Compute** the empirical reward function $\widehat{r}$ with (11) and empirical transition kernel $\widehat{P}_0$ with (10).
3: **for** $h=H,H-1,\ldots,1$ **do**
4: **Update** the robust Q-value estimate as (15) with $\beta_h(s,a,b,V)$ defined in (18).
5: **Compute** Nash policy for each $s\in\mathcal{S}$ as

$$\left(\mu_h^+(s),\nu_h^+(s)\right)=\mathsf{ComputNash}\left(\widehat{Q}_h^+(s,\cdot,\cdot)\right);$$
$$\left(\mu_h^-(s),\nu_h^-(s)\right)=\mathsf{ComputNash}\left(\widehat{Q}_h^-(s,\cdot,\cdot)\right).$$

6: **Update** the robust value estimate for each $s\in\mathcal{S}$ as

$$\widehat{V}_h^-(s)=\mathbb{E}_{a\sim\mu_h^-(s),b\sim\nu_h^-(s)}\left[\widehat{Q}_h^-(s,a,b)\right];\ \widehat{V}_h^+(s)=\mathbb{E}_{a\sim\mu_h^+(s),b\sim\nu_h^+(s)}\left[\widehat{Q}_h^+(s,a,b)\right].$$

7: **end for**
output The policy pair $(\widehat{\mu},\widehat{\nu})$, where $\widehat{\mu}=\{\mu_h^-\}_{h=1}^H$ and $\widehat{\nu}=\{\nu_h^+\}_{h=1}^H$.

Definition 4.1 (Robust unilateral clipped concentrability)**.** *We define* $C_{\mathrm{r}}^\star\in\left[\frac{1}{S(A+B)},\infty\right]$ *as the smallest value that satisfies*

$$\max\left\{\sup_{(\mu,s,a,b,h,P)\in\Delta(\mathcal{A})\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]\times\mathcal{U}^{\sigma^-}(P^0)}\frac{\min\left\{d_h^{\mu,\nu^\star,P}(s,a,b),\frac{1}{S(A+B)}\right\}}{d_h^{\mathsf{n},P^0}(s,a,b)},\right.$$
$$\left.\sup_{(\nu,s,a,b,h,P)\in\Delta(\mathcal{B})\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]\times\mathcal{U}^{\sigma^+}(P^0)}\frac{\min\left\{d_h^{\mu^\star,\nu,P}(s,a,b),\frac{1}{S(A+B)}\right\}}{d_h^{\mathsf{n},P^0}(s,a,b)}\right\}\le C_{\mathrm{r}}^\star \quad (20)$$

for the behavior policies of the historical dataset $\mathcal{D}$ *satisfies, and refer to it as the robust unilateral clipped concentrability coefficient. For consistency, we define the convention "*$0/0=0$*".*

Notably, if $d_h^{\mu,\nu^\star,P}(s,a,b)$ or $d_h^{\mu^\star,\nu,P}(s,a,b)$ is larger than $1/(S(A+B))$, the robust unilateral clipped concentrability assumption does not require the data distribution $d_h^{\mathsf{n},P^0}(s,a,b)$ to scale with $d_h^{\mu,\nu^\star,P}(s,a,b)$ or $d_h^{\mu^\star,\nu,P}(s,a,b)$ proportionally. Estimating the concentrability coefficient $C_{\mathrm{r}}^\star$ from offline data is generally information-theoretically impossible, as demonstrated even in the single-agent case by the example construction in Section 3.4 of [24]. Fortunately, this does not affect the execution of our algorithm.

Next, we outline the principal theoretical findings concerning the sample complexity of learning robust NE in RTZMGs, including an upper bound for RTZ-VI-LCB (Algorithm 2) and an information-theoretic lower bound. We start with the finite-sample guarantee for RTZ-VI-LCB, with the proof provided in Appendix D.

Theorem 4.2 (Upper bound for RTZ-VI-LCB)**.** *Under the TV uncertainty set* $\mathcal{U}^{\sigma^+}(\cdot)$ *and* $\mathcal{U}^{\sigma^-}(\cdot)$ *defined in (2) with* $\sigma^+,\ \sigma^-\in(0,1]$*, define* $d_{\mathsf{m}}^{\mathsf{n}}=\min_{h,s,a,b}\{d_h^{\mathsf{n}}(s,a,b):d_h^{\mathsf{n}}(s,a,b)>0\}$*, and let* $f(\sigma^+,\sigma^-)=\min\left\{(H\sigma^+-1+(1-\sigma^+)^H)/(\sigma^+)^2,(H\sigma^--1+(1-\sigma^-)^H)/(\sigma^-)^2,H\right\}$*. Consider any* $\delta\in(0,1)$ *and any RTZMG* $\mathcal{MG}_{\mathsf{r}}=\left\{\mathcal{S},\mathcal{A},\mathcal{B},\mathcal{U}^{\sigma^+}(P^0),\mathcal{U}^{\sigma^-}(P^0),r,H\right\}$*. For sufficient large constants* $c_0,c_1>0$*, with probability of at least* $1-\delta$*, we can achieve*

$$\mathrm{Gap}(\widehat{\mu},\widehat{\nu})\le c_1\sqrt{\frac{C_{\mathsf{r}}^\star H^3S(A+B)\log\frac{KH}{\delta}}{K}f(\sigma^+,\sigma^-,H)} \quad (21)$$

with the total number of samples T *exceeding*

$$T=KH\ge c_0\frac{H^2S(A+B)}{d_{\mathsf{m}}^{\mathsf{n}}}\log\frac{KH}{\delta}f(\sigma^+,\sigma^-,H). \quad (22)$$

Remark 4.3. Under a generative model with uniform sampling, the visitation distribution becomes explicit; that is, $d_h^{\mathsf{n},P^0}(s,a,b) = \frac{1}{SAB}$. In this case, it is sufficient to choose $C_{\mathrm{r}}^\star = \min\{A,B\}$, which can be readily verified to satisfy Eq. (20). This choice leads to the sample complexity $\widetilde{O}\left(\frac{H^4SAB}{\varepsilon^2}\cdot f(\sigma^+,\sigma^-,H)\right)$, which matches the bound established in [34] and confirms the efficiency of our approach in the generative model setting.

We derive an information-theoretic lower bound of sample complexity, with its proof in Appendix E.

Theorem 4.4 (Lower bound for RTZMGs)**.** *Consider any tuple* $\mathcal{MG}_{\mathsf{r}} = \left\{\mathcal{S},\mathcal{A},\mathcal{B},\mathcal{U}^{\sigma^+}(P^0),\mathcal{U}^{\sigma^-}(P^0),r,H\right\}$ *obeying* $H > 16\log 2$ *and* $\sigma^+,\ \sigma^- \in (0,1-c_0]$ *with any small efficiently positive constant* $0 < c_0 \le \frac{1}{4}$. *Let*

$$\varepsilon \le \begin{cases} \frac{c_2}{H}, & \text{if } \max\{\sigma^+,\sigma^-\} \le \frac{c_2}{2H}; \\ 1, & \text{otherwise}, \end{cases} \tag{23}$$

for any $c_2 \le \frac{1}{4}$. *With an initial state distribution* ϱ, *we can construct a set of RTZMGs* $\left\{\mathcal{M}_f^\phi | f \in \mathcal{F} = \{0,1,\cdots,SA-1\},\ \phi = [\phi_h]_{1\le h\le H} \in \Phi \subseteq \{0,1\}^H\right\}$. *For any dataset with* K *independent sample trajectories of length* H *per trajectory satisfying* $C \le C_{\mathrm{r}}^\star \le 2C$, *we have*

$$\inf_{\widehat{\mu},\widehat{\nu}} \max_{(f,\phi)\in\mathcal{F}\times\Phi} \left\{\mathbb{P}_\phi\big(\mathrm{Gap}(\widehat{\mu},\widehat{\nu}) > \varepsilon\big)\right\} \ge 1/8, \tag{24}$$

provided that

$$T = KH \le \frac{cC_{\mathrm{r}}^\star H^3 S(A+B)\min\left\{\frac{1}{\min\{\sigma^+,\sigma^-\}},H\right\}}{\varepsilon^2}. \tag{25}$$

Here, c *is an efficiently small constant. The infimum is obtained over all estimators* $(\widehat{\mu},\widehat{\nu})$.

These theorems offer the following key implications:

(i) **Theorem 4.2** demonstrates that the proposed RTZ-VI-LCB algorithm can attain an ε-robust NE solution when sample size exceeds $\widetilde{O}\left(\frac{C_{\mathrm{r}}^\star H^4 S(A+B)}{\varepsilon^2} f(\sigma^+,\sigma^-,H)\right)$, suggesting that the sample efficiency for robust offline TZMGs is strongly influenced by the dataset quality (quantified by $C_{\mathrm{r}}^\star$) and problem structure of RTZMGs (reflected in the occupancy distributions $d_{\mathsf{m}}^{\mathsf{n}}$). If $C_{\mathrm{r}}^\star$ is as small as $\frac{1}{S(A+B)}$, the upper bound of the sample complexity exhibits a weaker dependency on actions $\{A,B\}$ and state S. Combining this upper bound with the lower bound in Theorem 4.4 shows that RTZ-VI-LCB's sample complexity is optimal w.r.t. key factors S, A, B and ε. This is the first optimal sample complexity upper bound for offline RTZMGs, regarding state S and actions $\{A,B\}$.

(ii) **Theorem 4.4** conveys two important points. When the uncertainty level is small (i.e., $\min\{\sigma^+,\sigma^-\} \lesssim \frac{1}{H}$), no algorithm can find an ε-optimal robust policy with fewer than $\Omega\left(\frac{C_{\mathrm{r}}^\star SH^4(A+B)}{\varepsilon^2}\right)$ samples for all offline RTZMGs, matching the complexity requirement for non-robust offline TZMGs [18]. This implies that robust TZMGs are at least as challenging as standard TZMGs for low uncertainty. When the uncertainty level satisfies $\min\{\sigma^+,\sigma^-\} \gtrsim \frac{1}{H}$, no algorithm can find an ε-optimal robust policy with the numbers of samples fewer than $\Omega\left(\frac{C_{\mathrm{r}}^\star SH^3(A+B)}{\varepsilon^2\min\{\sigma^+,\sigma^-\}}\right)$. To this end, RTZ-VI-LCB is the first provably optimal algorithm on S and $\{A,B\}$ for RTZMGs without requiring full coverage assumptions.

Moreover, our algorithm can be extended to multi-player general-sum MGs with m players and A_i actions and uncertainty level σ_i per player; see Appendix F. We can obtain the following theoretical guarantee for this extended algorithm, named Multi-RTZ-VI-LCB:

Theorem 4.5 (Upper bound for Multi-RTZ-VI-LCB)**.** *Consider any* $\delta \in (0,1)$ *and any robust multi-player general-sum MGs* $\mathcal{MG}_{\mathsf{r}} = \mathcal{M}(\mathcal{S},\{\mathcal{A}_i\}_{i=1}^m,H,\{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{i=1}^m,\{r_i\}_{i=1}^m)$. *Under the TV uncertainty set* $\mathcal{U}^{\sigma_i}(\cdot)$ *defined in (2) with* $\sigma_i \in (0,1]$ *for* $i = 1,2,\cdots,m$. *Define* $d_{\mathsf{m}}^{\mathsf{n}} = \min_{h,s,\boldsymbol{a}}\{d_h^{\mathsf{n}}(s,\boldsymbol{a}) : d_h^{\mathsf{n}}(s,\boldsymbol{a}) > 0\}$, *and* $f(\{\sigma_i\}_{i=1}^m,H) = \min\left\{\left\{\frac{(H\sigma_i-1+(1-\sigma_i)^H)}{(\sigma_i)^2}\right\}_{i=1}^m, H\right\}$. *For sufficient large constants* $c_0,c_1 > 0$, *with probability of at least* $1-\delta$, *we can achieve*

$$\mathrm{Gap}(\widehat{\pi}) \le c_1\sqrt{\frac{C_{\mathrm{r}}^\star H^3 S\sum_{i=1}^m A_i}{K}\log\frac{KH}{\delta}f(\{\sigma_i\}_{i=1}^m,H)},$$

with the total number of samples T exceeding

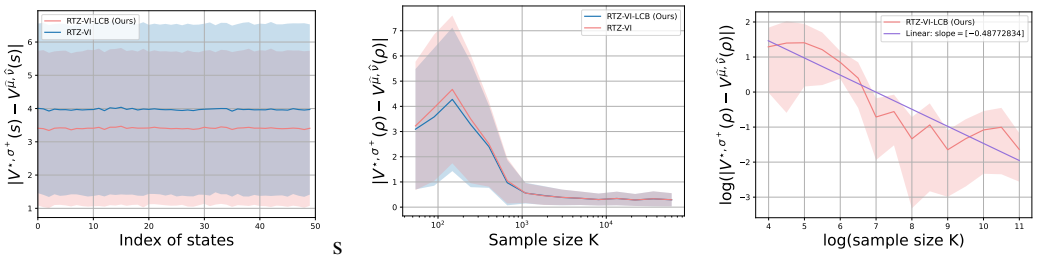
$$T = KH \geq c_0 \frac{H^2 S \sum_{i=1}^m A_i}{d_{\mathsf{m}}^{\mathsf{n}}} \log \frac{KH}{\delta} f(\{\sigma_i\}_{i=1}^m, H). \tag{26}$$

Theorem 4.5 demonstrates that Multi-RTZ-VI-LCB can attain an ε-robust NE solution when the sample size exceeds $\widetilde{O}\left(\frac{C_{\mathsf{r}}^\star H^4 S \sum_{i=1}^m A_i}{\varepsilon^2} f(\{\sigma_i\}_{i=1}^m, H)\right)$, breaking the curse of multiagency.

5 Numerical Experiments

To effectively evaluate our algorithm, we have conducted numerical experiments on randomly generated transition kernels, following the code proposed by [39]. In particular, we adopt the parameter setting as $S = 50$, $A = B = 2$, and $H = 100$, averaged over 100 seeds. Experiments are conducted on PyTorch 2.0.0 with a single NVIDIA RTX 4090 24GB GPU. In our experiments, the robust NE at each state and timestep is computed using standard NE solvers, i.e., the Python package nashpy. Our algorithm is compatible with any exact or approximate NE solver, including computational relaxations or sampling-based methods.

As shown in Figure 1(a), the case of $K = 148 \approx e^5$ demonstrates that our proposed algorithm consistently outperforms the baseline value iteration for robust TZMGs (RTZ-VI) across all states and all sample sizes. This trend remains consistent across other values of K as well. Moreover, we have plotted the sub-optimality performance gap of RTZ-VI-LCB w.r.t. the sample size on a log-log scale to corroborate the scaling of the sample size on the performance gap. Fitting using linear regression leads to a slope estimate of -0.4877. This nicely matches the finding of our theoretical guarantee.



(a) Value gap versus states (b) Value gap versus sample size (c) Value gap dependency w.r.t. K

Figure 1: The performances of RTZ-VI-LCB and RTZ-VI in the stochastic TZMG problem.

6 Conclusion

In this paper, we design an efficient robust model-based algorithm for offline RTZMGs, which is value iteration with lower confidence bounds for RTZMGs. Our algorithm integrates robust value iteration with the principle of pessimism. By imposing a tailored assumption (robust unilateral clipped concentrability) on the historical dataset to account for the distribution shift, we address robustness in the worse-case scenario of the shared environment, analyze the finite-sample complexity of the proposed RTZ-VI-LCB algorithm, and establish an information-theoretic lower bound to evaluate its optimality across various uncertainty levels. To the best of our knowledge, this is the first provably optimal algorithm for offline RTZMGs that addresses the dependency on states S and actions $\{A, B\}$, while accounting for model perturbations and partial coverage. Furthermore, we extend RTZ-VI-LCB to multi-agent general-sum MGs, demonstrating a breakthrough in breaking the curse of multiagency.

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

Acknowledgements

The work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants U21B2029, 62531019, and U21A20456; in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010006; and in part by a Nokia Donation Project. The work of Na Li was supported by the China Scholarship Council.

References

[1] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[2] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

[3] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167:235–292, 2018.

[4] Sushrut Bhalla, Sriram Ganapathi Subramanian, and Mark Crowley. Deep multi agent reinforcement learning for autonomous driving. In *Canadian Conference on Artificial Intelligence*, pages 67–78. Springer, 2020.

[5] Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[7] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

[8] Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022.

[9] Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

[10] Qiwen Cui, Kaiqing Zhang, and Simon Du. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2651–2652. PMLR, 2023.

[11] Constantinos Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

[12] John C Duchi. Introductory lectures on stochastic optimization. *The mathematics of data*, 25:99–186, 2018.

[13] Songtao Feng, Ming Yin, Yu-Xiang Wang, Jing Yang, and Yingbin Liang. Model-free algorithm with improved sample efficiency for zero-sum markov games, 2023.

[14] Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

[15] Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

[16] Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

[17] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[18] Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent RL. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*, 2022.

[19] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[20] Shyam Sundar Kannan, Vishnunandan LN Venkatesh, and Byung-Cheol Min. Smart-llm: Smart multi-agent robot task planning using large language models. *arXiv preprint arXiv:2309.10062*, 2023.

[21] Erim Kardeş, Fernando Ordóñez, and Randolph W Hall. Discounted robust stochastic games and an application to queueing control. *Operations research*, 59(2):365–382, 2011.

[22] Nathan Lambert, Markus Wulfmeier, William Whitney, Arunkumar Byravan, Michael Bloesch, Vibhavari Dasagi, Tim Hertweck, and Martin Riedmiller. The challenges of exploration for offline reinforcement learning. *arXiv preprint arXiv:2201.11861*, 2022.

[23] Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Linear last-iterate convergence for matrix games and stochastic games. *arXiv preprint arXiv:2006.09517*, 2020.

[24] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

[25] Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024.

[26] Na Li, Yuchen Jiao, Hangguan Shan, and Shefeng Yan. Provable memory efficient self-play algorithm for model-free reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[27] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[28] Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7001–7010. PMLR, July 2021.

[29] Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.

[30] Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

[31] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10199–10210, 2020.

[32] Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

[33] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Forty-first International Conference on Machine Learning*, 2024.

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

[36] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[37] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[38] Daniel Vial, Sanjay Shakkottai, and R Srikant. Robust multi-agent bandits over undirected graphs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(3):1–57, 2022.

[39] Antony Vijesh et al. A multi-step minimax q-learning algorithm for two-player zero-sum markov games. *arXiv preprint arXiv:2407.04240*, 2024.

[40] Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2793–2848. PMLR, 2023.

[41] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. *Math. Oper. Res.*, 48(1):433–462, Jun. 2022.

[42] Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9728–9754. PMLR, 2023.

[43] Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. Model-based reinforcement learning for offline zero-sum markov games. *Operations Research*, 2024.

[44] Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

[45] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021.

[46] Lanting Zeng, Dawei Qiu, and Mingyang Sun. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. *Applied Energy*, 324:119688, 2022.

[47] Huan Zhang, Hongge Chen, Duane S Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations*, 2021.

[48] Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33:10571–10583, 2020.

[49] Ziyuan Zhou and Guanjun Liu. Robustness testing for multi-agent reinforcement learning: State perturbations on critical agents. *arXiv preprint arXiv:2306.06136*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction precisely reflect the contribution and scope of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We explicitly list all key Theorems in Section 4, and provide complete proofs in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed illustration of our proposed algorithm in the Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is available at `https://github.com/NLee10/RTZ-VI-LCB`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide full details in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the average results across 100 seeds with standard deviations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We point out the specific compute resources in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper obeys the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: There is no societal impact of the work performed since this work focuses on theoretical analysis.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We respect the Licenses for existing assets that we use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We will release new assets proposed in our paper once the paper is accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper does not use LLM to impact the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Contents

A Numerical Experiments

$\log(KH)$	6	7	8	9
RTZ-VI-LCB	**3.1699**	**2.1498**	**0.2324**	**0.0819**
$\mathrm{P^2M^2PO}$	3.5970	2.2503	0.2404	0.0890

These results highlight the superiority of our method across varying dataset sizes and validate the theoretical claims under controlled experimental conditions. While we focus on tabular settings in this work, our algorithm can serve as a theoretical foundation for scalable extensions based on linear or kernel-based function approximation.

B Preliminaries

In this appendix, we introduce the dual equivalence of robust Bellman operators and key facts about RTZMGs and their empirical counterparts, laying the foundation for our theoretical analysis.

B.1 Dual equivalence of robust Bellman

We can compute the robust Bellman operator by solving its dual formulation rather than the original form, as long as the predefined uncertainty set is in a benign form (e.g., utilizing TV distance as the divergence function) [17, 33]. Taking TV distance as an example, we describe the equivalence under strong duality between the robust Bellman operator and its dual form as Lemma B.1.

Lemma B.1. *Consider any TV uncertainty set $\mathcal{U}^{\sigma^+}(P)$ and $\mathcal{U}^{\sigma^-}(P)$ associated with fixed uncertainty levels $\sigma^+, \sigma^- \in (0,1]$ and any probability vector $P \in \Delta(\mathcal{S})$. For any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, one has*

$$\inf_{P \in \mathcal{U}^{\sigma^+}(P)} PV = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P[V]_\alpha - \sigma^+ \left(\alpha - \min_{s'} [V]_\alpha (s') \right) \right\}; \tag{27a}$$

$$\sup_{P \in \mathcal{U}^{\sigma^-}(P)} PV = \min_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P[V]_\alpha - \sigma^- \left(\alpha - \max_{s'} [V]_\alpha (s') \right) \right\}, \tag{27b}$$

where $[V]_\alpha$ is defined in (17)

Lemma B.1 can be proved similarly to Lemma 4.3 in [17]. Compared the standard Bellman operator, this lemma guarantees that no additional computing cost is required when applying the robust Bellman operator.

B.2 Facts of RTZMGs and empirical RTZMGs

Recall the definition of any RTZMG $\mathcal{MG}_{\mathsf{r}} = \left\{ \mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{U}_\rho^{\sigma^+}(P^0), \mathcal{U}_\rho^{\sigma^-}(P^0), r, H \right\}$. According to robust Bellman equations in (7), one has: for any product policy (μ, ν) and any $(h, s, a, b) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B}$,

$$Q_h^{\mu,\nu,\sigma^+}(s,a,b) = r_h(s,a,b) + \inf_{P \in \mathcal{U}_\rho^{\sigma^+}(P^0_{h,s,a,b})} PV_{h+1}^{\mu,\nu,\sigma^+}; \tag{28a}$$

$$Q_h^{\mu,\nu,\sigma^-}(s,a,b) = r_h(s,a,b) + \sup_{P \in \mathcal{U}_\rho^{\sigma^-}(P^0_{h,s,a,b})} PV_{h+1}^{\mu,\nu,\sigma^-}, \tag{28b}$$

where

$$V_h^{\mu,\nu,\sigma^+}(s) = \mathbb{E}_{a \sim \mu_h(s), b \sim \nu_h(s)} \left[Q_h^{\mu,\nu,\sigma^+}(s,a,b) \right];$$
$$V_h^{\mu,\nu,\sigma^-}(s) = \mathbb{E}_{a \sim \mu_h(s), b \sim \nu_h(s)} \left[Q_h^{\mu,\nu,\sigma^-}(s,a,b) \right].$$

Considering the offline setting, we use $\widehat{\mathcal{MG}}_{\mathsf{r}} = \left\{ \mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{U}_\rho^{\sigma^+}(\widehat{P}^0), \mathcal{U}_\rho^{\sigma^-}(\widehat{P}^0), r, H \right\}$ to represent the empirical RTZMG, which is established along with the estimated nominal distribution $\widehat{P}^0$

in (10). Therefore, for any product policy (μ, ν), we define the empirical robust value function (resp. empirical robust Q-function) in $\widehat{\mathcal{MG}}_{\mathsf{r}}$ as $\widehat{V}_h^{\mu,\nu,\sigma^+}$ and $\widehat{V}_h^{\mu,\nu,\sigma^-}$ (resp. $\widehat{Q}_h^{\mu,\nu,\sigma^+}$ and $\widehat{Q}_h^{\mu,\nu,\sigma^-}$), which are analogous to (6). Moreover, we can similarly define the optimal empirical robust value function for both players over $\widehat{\mathcal{MG}}_{\mathsf{r}}$: $\forall s \in \mathcal{S}$,

$$\widehat{V}_h^{\star,\nu,\sigma^+}(s) = \widehat{V}_h^{\mu^\star,\nu,\sigma^+}(s) \coloneqq \max_{\mu:\mathcal{S}\times[H]\to\Delta(\mathcal{A})} \widehat{V}_h^{\mu,\nu,\sigma^+}(s) = \max_{\mu:\mathcal{S}\times[H]\to\Delta(\mathcal{A})} \inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}^0)} \widehat{V}_h^{\mu,\nu,P}(s); \tag{29a}$$

$$\widehat{V}_h^{\mu,\star,\sigma^-}(s) = \widehat{V}_h^{\mu,\nu^\star,\sigma^-}(s) \coloneqq \max_{\nu:\mathcal{S}\times[H]\to\Delta(\mathcal{B})} \widehat{V}_h^{\mu,\nu,\sigma^-}(s) = \max_{\nu:\mathcal{S}\times[H]\to\Delta(\mathcal{B})} \inf_{P\in\mathcal{U}^{\sigma^-}(\widehat{P}^0)} \widehat{V}_h^{\mu,\nu,P}(s). \tag{29b}$$

Notably, for all $s \in \mathcal{S}$, there exists at least one *robust best-response* policy that can achieve $\widehat{V}_h^{\star,\nu,\sigma^+}(s)$ and $\widehat{V}_h^{\mu,\star,\sigma^-}(s)$, as proved in [5]. Therefore, we can obtain the empirical robust Bellman equation similar to (7) as: for any product policy (μ, ν),

$$\widehat{Q}_h^{\mu,\nu,\sigma^+}(s,a,b) = r_h(s,a,b) + \inf_{P\in\mathcal{U}_\rho^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} P\widehat{V}_{h+1}^{\mu,\nu,\sigma^+}; \tag{30a}$$

$$\widehat{Q}_h^{\mu,\nu,\sigma^-}(s,a,b) = r_h(s,a,b) + \sup_{P\in\mathcal{U}_\rho^{\sigma^-}(\widehat{P}^0_{h,s,a,b})} P\widehat{V}_{h+1}^{\mu,\nu,\sigma^-}, \tag{30b}$$

where

$$\widehat{V}_h^{\mu,\nu,\sigma^+}(s) = \mathbb{E}_{a\sim\mu_h(s),b\sim\nu_h(s)}[\widehat{Q}_h^{\mu,\nu,\sigma^+}(s,a,b)];$$
$$\widehat{V}_h^{\mu,\nu,\sigma^-}(s) = \mathbb{E}_{a\sim\mu_h(s),b\sim\nu_h(s)}[\widehat{Q}_h^{\mu,\nu,\sigma^-}(s,a,b)].$$

C Proof of Lemma 3.1

C.1 Independence property

Let us examine two distinct data-generation mechanisms, where a sample transition quadruple (s, a, b, h, s') represents a transition from state s with actions (a, b) to state s' at step h.

Step 1: Creating $\mathcal{D}^{\mathrm{t,a}}$ based on $\mathcal{D}^{\mathrm{t}}$. To construct the augmented dataset $\mathcal{D}^{\mathrm{t,a}}$, for each $(s, h) \in \mathcal{S} \times [H]$, we first include all $N_h^{\mathrm{t}}(s)$ sample transitions in $\mathcal{D}^{\mathrm{t}}$ originating from state s at step h in $\mathcal{D}^{\mathrm{t,a}}$. If $N_h^{\mathrm{t}}(s) > N_h^{\mathrm{m}}(s)$, we supplement $\mathcal{D}^{\mathrm{t,a}}$ with additional $[N_h^{\mathrm{t}}(s) - N_h^{\mathrm{m}}(s)]$ independent sample transitions $\big\{\big(s, a_{h,s}^{(i)}, b_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big)\big\}$, as follows:

$$a_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \mu_h^{\mathrm{b}}(\cdot|s), \quad b_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \nu_h^{\mathrm{b}}(\cdot|s), \quad s_{h,s}^{\prime(i)} \overset{\text{i.i.d.}}{\sim} P_h\big(\,\cdot\,|s, a_{h,s}^{(i)}, b_{h,s}^{(i)}\big), \quad N_h^{\mathrm{m}}(s) < i \le N_h^{\mathrm{t}}(s).$$

Step 2: Constructing $\mathcal{D}^{\mathrm{iid}}$. For each $(s, h) \in \mathcal{S} \times [H]$, we generate $N_h^{\mathrm{t}}(s)$ independent sample transitions $\big\{\big(s, a_{h,s}^{(i)}, b_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big)\big\}$, as follows:

$$a_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \mu_h^{\mathrm{b}}(\cdot|s), \quad b_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \nu_h^{\mathrm{b}}(\cdot|s), \quad s_{h,s}^{\prime(i)} \overset{\text{i.i.d.}}{\sim} P_h\big(\,\cdot\,|s, a, b\big), \quad 1 \le i \le N_h^{\mathrm{t}}(s).$$

The resulting dataset is defined as:

$$\mathcal{D}^{\mathrm{iid}} \coloneqq \Big\{\big(s, a_{h,s}^{(i)}, b_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big) \mid s \in \mathcal{S}, 1 \le h \le H, 1 \le i \le N_h^{\mathrm{t}}(s)\Big\}.$$

Establishing independence property. The dataset $\mathcal{D}^{\mathrm{t,a}}$ deviates from $\mathcal{D}^{\mathrm{t}}$ only if $N_h^{\mathrm{t}}(s) > N_h^{\mathrm{m}}(s)$. This augmentation ensures that $\mathcal{D}^{\mathrm{t,a}}$ contains precisely $N_h^{\mathrm{t}}(s)$ sample transitions from state s at step h. Both $\mathcal{D}^{\mathrm{t,a}}$ and $\mathcal{D}^{\mathrm{iid}}$ comprise exactly $N_h^{\mathrm{t}}(s)$ sample transitions from state s at step h, with $\{N_h^{\mathrm{t}}(s)\}$ being statistically independent of random sample generation. Consequently, given $\{N_h^{\mathrm{t}}(s)\}$, the sample transitions in $\mathcal{D}^{\mathrm{t,a}}$ across different steps are statistically independent. As a result, both $\mathcal{D}^{\mathrm{t}}$ and $\mathcal{D}^{\mathrm{iid}}$ can be regarded as collections of independent samples.

C.2 Proof of $N_h^{\mathsf{t}}(s) \le N_h^{\mathsf{m}}(s)$

Since $\mathcal{D}^{\mathsf{a}}$ is generated by half of the sample trajectories in line 2 in Algorithm 1, we have

$$N_h^{\mathsf{a}}(s) = \sum_{k=K/2+1}^{K} \mathbb{1}\{s_h^k = s\},\ \forall s \in \mathcal{S}, 1 \le h \le H.$$

Thus, we can view $N_h^{\mathsf{a}}(s)$ as the sum of $K/2$ independent Bernoulli random variables with mean $d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)$. According to the Bernstein inequality and the union bound, we derive

$$\begin{aligned}
&\mathbb{P}\left\{\exists (s,h) \in \mathcal{S}\times[H] : \left|N_h^{\mathsf{a}}(s) - \frac{K}{2} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\right| \ge N_0\right\} \\
\le &\sum_{s\in\mathcal{S}, h\in[H]} \mathbb{P}\left\{\left|N_h^{\mathsf{a}}(s) - \frac{K}{2} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\right| \ge N_0\right\} \\
\le &2HS\exp\left(-\frac{N_0^2/2}{N_{h,s} + N_0/3}\right), \quad \forall N_0 \ge 0,
\end{aligned}$$

where

$$N_{h,s} \coloneqq \frac{K}{2}\mathsf{Var}\big(\mathbb{1}\{s_h^t = s\}\big) = \frac{K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\big(1 - d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\big)}{2} \le \frac{K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)}{2}.$$

Therefore, with probability of at least $1-2\delta$, we have that: $\forall s \in \mathcal{S}$ and $\forall 1 \le h \le H$,

$$\left|N_h^{\mathsf{a}}(s) - \frac{K}{2} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\right| \le \sqrt{4N_{h,s}\log\frac{HS}{\delta}} + \frac{2}{3}\log\frac{HS}{\delta} \le \sqrt{2K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\log\frac{HS}{\delta}} + \log\frac{HS}{\delta}. \tag{31}$$

Since $\mathcal{D}^{\mathsf{m}}$ and $\mathcal{D}^{\mathsf{a}}$ are generated in the same way, we obtain that with probability exceeding $1-2\delta$,

$$\left|N_h^{\mathsf{m}}(s) - \frac{K}{2} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\right| \le \sqrt{2K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\log\frac{HS}{\delta}} + \log\frac{HS}{\delta},\ \forall s \in \mathcal{S}, 1 \le h \le H. \tag{32}$$

By combining (31) and (32), it follows that

$$|N_h^{\mathsf{m}}(s) - N_h^{\mathsf{a}}(s)| \le 2\sqrt{2K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\log\frac{HS}{\delta}} + 2\log\frac{HS}{\delta},\ \forall s \in \mathcal{S}, 1 \le h \le H. \tag{33}$$

Now, we prove $N_h^{\mathsf{t}}(s) \le N_h^{\mathsf{m}}(s)$ by considering two cases. *In the first case,* $N_h^{\mathsf{a}}(s) \le 100\log\frac{HS}{\delta}$. According to the definition in (13), we obtain

$$N_h^{\mathsf{t}}(s) = \max\left\{N_h^{\mathsf{a}}(s) - 10\sqrt{N_h^{\mathsf{a}}(s)\log\frac{HS}{\delta}},\ 0\right\} = 0 \le N_h^{\mathsf{m}}(s). \tag{34}$$

In the second case, $N_h^{\mathsf{a}}(s) > 100\log\frac{HS}{\delta}$. From (31), it follows that

$$\frac{K}{2} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) + \sqrt{2K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\log\frac{HS}{\delta}} + \log\frac{HS}{\delta} \ge N_h^{\mathsf{a}}(s),$$

which leads to

$$K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) \ge (9\sqrt{2})^2 \log\frac{HS}{\delta} \ge 100\log\frac{HS}{\delta}. \tag{35}$$

By substituting (35) into (31), we obtain

$$N_h^{\mathsf{a}}(s) \ge \frac{K}{2} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) - \sqrt{2K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s)\log\frac{HS}{\delta}} - \log\frac{HS}{\delta} \ge \frac{K}{4} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s). \tag{36}$$

Consequently, in the case of $N_h^{\mathsf{a}}(s) > 100 \log \frac{HS}{\delta}$, we have

$$\begin{aligned} N_h^{\mathsf{t}}(s) &= \max\left\{ N_h^{\mathsf{a}}(s) - 10\sqrt{N_h^{\mathsf{a}}(s) \log \frac{HS}{\delta}},\, 0 \right\} \\ &= N_h^{\mathsf{a}}(s) - 10\sqrt{N_h^{\mathsf{a}}(s) \log \frac{HS}{\delta}} \\ &\overset{\text{(i)}}{\leq} N_h^{\mathsf{a}}(s) - 5\sqrt{K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) \log \frac{HS}{\delta}} \\ &\overset{\text{(ii)}}{\leq} N_h^{\mathsf{a}}(s) - \left\{ 2\sqrt{2K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) \log \frac{HS}{\delta}} + 2\log \frac{HS}{\delta} \right\} \overset{\text{(iii)}}{\leq} N_h^{\mathsf{m}}(s), \end{aligned} \tag{37}$$

where (i) holds under condition (36), (ii) holds under the condition (35), and (iii) comes from the inequality (33) with probability of at least $1 - 2\delta$.

Combining (34) and (37), we establish $N_h^{\mathsf{t}}(s) \leq N_h^{\mathsf{m}}(s)$.

As proved in Appendix C.3, we have: $\forall (s, a, b, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$, with probability exceeding $1 - 2\delta$,

$$N_h^{\mathsf{t}}(s, a, b) \geq N_h^{\mathsf{t}}(s) \mu_h^{\mathsf{n}}(a \,|\, s) \nu_h^{\mathsf{n}}(b \,|\, s) - \sqrt{4 N_h^{\mathsf{t}}(s) \mu_h^{\mathsf{n}}(a \,|\, s) \nu_h^{\mathsf{n}}(b \,|\, s) \log \frac{KH}{\delta}} - \log \frac{KH}{\delta}. \tag{38}$$

Employing the fact $N_h^{\mathsf{t}}(s) \leq N_h^{\mathsf{m}}(s)$ and (38), we can prove (14) in two cases, i.e., $K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) \leq 1600 \log \frac{KH}{\delta}$ and $K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) > 1600 \log \frac{KH}{\delta}$.

In the case of $K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a) \leq 1600 \log \frac{KH}{\delta}$, we can readily obtain

$$\frac{K}{8} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a) - 5\sqrt{K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a) \log \frac{KH}{\delta}} \leq 0 \leq N_h^{\mathsf{t}}(s, a). \tag{39}$$

In the case of $K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) = K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) \mu_h^{\mathsf{n}}(a \,|\, s) \nu_h^{\mathsf{n}}(b \,|\, s) > 1600 \log \frac{KH}{\delta}$, we obtain

$$N_h^{\mathsf{a}}(s) \geq \frac{K}{4} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) \geq 400 \log \frac{KH}{\delta}, \tag{40}$$

which can be derived following the same line of (36). Then, (40) and the definition of $N_h^{\mathsf{t}}(s)$ together yield

$$\begin{aligned} N_h^{\mathsf{t}}(s) &\geq N_h^{\mathsf{a}}(s) - 10\sqrt{N_h^{\mathsf{a}}(s) \log \frac{KH}{\delta}} \\ &\geq \frac{K}{4} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) - 10\sqrt{\frac{K}{4} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) \log \frac{KH}{\delta}} \geq \frac{K}{8} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s). \end{aligned}$$

As a consequent,

$$N_h^{\mathsf{t}}(s) \mu_h^{\mathsf{n}}(a \,|\, s) \nu_h^{\mathsf{n}}(b \,|\, s) \geq \frac{K}{8} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s) \mu_h^{\mathsf{n}}(a \,|\, s) \nu_h^{\mathsf{n}}(b \,|\, s) \tag{41}$$

$$= \frac{K}{8} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) \geq 200 \log \frac{KH}{\delta}, \tag{42}$$

where the last inequality holds under the assumption of the second case. Combining (41) with (38) yields

$$\begin{aligned} N_h^{\mathsf{t}}(s, a, b) &\geq \frac{K}{8} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) - \sqrt{\frac{K}{2} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) \log \frac{KH}{\delta}} - \log \frac{KH}{\delta} \\ &\geq \frac{K}{8} d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) - 2\sqrt{K d_h^{\mu^{\mathsf{n}},\nu^{\mathsf{n}}}(s, a, b) \log \frac{KH}{\delta}}. \end{aligned}$$

Combining the result above with (39) and referencing (38), we conclude the proof of Lemma 3.1.

C.3 Proof of (38).

To prove (38), we analyze two cases, i.e., $N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) \leq 4\log\frac{KH}{\delta}$ and $N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) > 4\log\frac{KH}{\delta}$.

In the first case of $N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) \leq 4\log\frac{KH}{\delta}$, we conclude the right-hand side of (38) is negative, leading to (38). In the second case of $N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) > 4\log\frac{KH}{\delta}$, we compose a special set $\mathcal{D}^{\mathsf{l}}$ as

$$\mathcal{D}^{\mathsf{l}} \coloneqq \left\{(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H] \;\middle|\; N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) > 4\log\frac{KH}{\delta}\right\}. \tag{43}$$

Given the fact that

$$\begin{aligned}\sum_{(s,a,b,h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]} N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) &= \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{t}}(s) \sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} \mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) \\ &= \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{t}}(s) \leq \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{a}}(s) = \frac{KH}{2},\end{aligned}$$

the cardinality of $\mathcal{D}^{\mathsf{l}}$ can be bounded as:

$$\left|\mathcal{D}^{\mathsf{l}}\right| < \frac{\sum_{(s,a,b,h)} N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s)}{4\log\frac{KH}{\delta}} \leq KH/2. \tag{44}$$

Moreover, we can view $N_h^{\mathsf{t}}(s,a)$ as the sum of $N_h^{\mathsf{t}}(s)$ independent Bernoulli random variables with mean $\mu_h^{\mathsf{n}}(a \mid s)\nu_h^{\mathsf{n}}(b \mid s)$, due to $N_h^{\mathsf{t}}(s) \leq N_h^{\mathsf{m}}(s)$ with high probability and conditioned on $N_h^{\mathsf{t}}(s)$ and $N_h^{\mathsf{m}}(s)$. Analogous to (31) based on the condition $N_h^{\mathsf{t}}(s) \leq N_h^{\mathsf{m}}(s)$, we can repeat the Bernstein-type argument and obtain that for any fixed triple (s,a,b,h), with probability of at least $1-2\delta/(KH)$,

$$N_h^{\mathsf{t}}(s,a,b) \geq N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s) - \sqrt{4N_h^{\mathsf{t}}(s)\mu_h^{\mathsf{n}}(a\,|\,s)\nu_h^{\mathsf{n}}(b\,|\,s)\log\frac{KH}{\delta}} - \log\frac{KH}{\delta}. \tag{45}$$

With probability exceeding $1-\delta$, (45) holds $\forall(s,a,b,h) \in \mathcal{D}^{\mathsf{l}}$ by utilizing the union bound of (44) over all $(s,a,b,h) \in \mathcal{D}^{\mathsf{l}}$. Combining the two cases, we conclude that (38) holds $\forall(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$ with probability of at least $1-\delta$.

D Proof of Theorem 4.2

Theorem 4.2 can be proved in three steps.

D.1 Step 1: Decoupling statistical dependency

Before bounding $\mathrm{Gap}(\widehat{\mu},\widehat{\nu})$, we introduce an important lemma, quantifying the difference between $\widehat{P}$ and P when projected in the direction of the value function.

Lemma D.1. *Instate the assumptions in Theorem 4.2. Consider any vector $V \in \mathbb{R}^S$ with $\|V\|_\infty \leq H$ for all $(h,s,a,b) \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ satisfying $N_h(s,a,b) > 0$. With probability of at least $1-\delta$, one has*

$$\left|\inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} PV - \inf_{P\in\mathcal{U}^{\sigma^+}(P^0_{h,s,a,b})} PV\right| \leq C_4\sqrt{\frac{1}{N_h(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V)\log\frac{KH}{\delta}} + C_4\frac{H\log\frac{KH}{\delta}}{N_h(s,a,b)} \tag{46}$$

for some sufficiently large constant $C_4 > 0$, and

$$\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) \leq 2\mathsf{Var}_{P^0_{h,s,a,b}}(V) + O\left(\frac{H^2}{N_h(s,a,b)}\log\frac{KH}{\delta}\right). \tag{47}$$

Proof can be found in Appendix G.1.

In simple terms, (46) provides a Bernstein-type concentration bound, while (47) ensures that the empirical variance estimate (i.e., the plug-in estimate) closely matches the true variance. Notably, Lemma D.1 does not require V to be statistically independent of $\widehat{P}^0_{h,s,a,b}$, which is essential given the complex statistical dependencies in RTZ-VI-LCB. Under the leave-one-out analysis (see, e.g., [1, 7, 24, 25]), we prove Lemma D.1 to decouple statistical dependencies, as illustrated in Appendix G.1.

With Lemma D.1, we now have: for any $(h,s,a,b) \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ satisfying $N_h(s,a,b) \geq 1$,

$$\left| \inf_{\mathcal{P} \in \mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} PV - \inf_{\mathcal{P} \in \mathcal{U}^{\sigma^+}(P^0_{h,s,a,b})} PV \right| \leq \beta_h\left(s,a,b,V\right). \tag{48}$$

Therefore, we conclude that $\widehat{Q}^+_h(s,a,b)$ is an optimistic estimation of $\widehat{Q}^{\mu,\nu,\sigma^+}_h(s,a,b)$, as summarized below.

Lemma D.2. *With probability exceeding $1-\delta$, it holds that*

$$\widehat{Q}^+_h(s,a,b) \geq Q^{\star,\widehat{\nu},\sigma^+}_h(s,a,b) \qquad \textit{and} \qquad \widehat{V}^+_h(s) \geq V^{\star,\widehat{\nu},\sigma^+}_h(s); \tag{49}$$

See Appendix G.2 for detailed proofs.

Moreover, we introduce another key lemma highlighting the difference between RTZMGs and standard TZMGs from the same idea of Lemma 3 in [34]. The range of the robust value function narrows as the uncertainty level σ^+ of its uncertainty set increases, as shown below.

Lemma D.3. *Consider the uncertainty set $\mathcal{U}^{\sigma^+}(\cdot)$ with TV distance and any RTZMG $\mathcal{MG}_{\mathsf{r}} = \left\{\mathcal{S},\mathcal{A},\mathcal{B},\mathcal{U}^{\sigma^+}(P),\mathcal{U}^{\sigma^-}(P),r,H\right\}$. The optimistic robust value function estimate $\widehat{V}^+_h$:*

$$\forall h \in [H]: \quad \max_{s\in\mathcal{S}} \widehat{V}^+_h - \min_{s\in\mathcal{S}} \widehat{V}^+_h \leq \min\left\{ \frac{(H+1)\left(1-(1-\sigma^+)^{H-h}\right)}{\sigma^+}, H \right\}.$$

See Appendix G.3 for detail proofs.

D.2 Step 2: Decomposing the error $\mathrm{Gap}(\widehat{\mu},\widehat{\nu})$

The goal of RTZ-VI-LCB is to output an ε-robust NE policy $(\widehat{\mu},\widehat{\nu})$ satisfying $\mathrm{Gap}(\widehat{\mu},\widehat{\nu})$ in (9), i.e.,

$$\mathrm{Gap}(\widehat{\mu},\widehat{\nu}) \coloneqq \max\left\{ V^{\star,\widehat{\nu},\sigma^+}_1(\varrho) - V^{\star,\sigma^+}_1(\varrho),\ V^{\star,\sigma^-}_1(\varrho) - V^{\widehat{\mu},\star,\sigma^-}_1(\varrho) \right\} \leq \varepsilon.$$

Due to the interchangeability between the max-player and the min-player, we assume without loss of generality that $V^{\star,\widehat{\nu},\sigma^+}_1(\varrho) - V^{\star,\sigma^+}_1(\varrho)$ is larger than $V^{\star,\sigma^-}_1(\varrho) - V^{\widehat{\mu},\star,\sigma^-}_1(\varrho)$, leading to $\mathrm{Gap}(\widehat{\mu},\widehat{\nu}) \leq \left\{ V^{\star,\widehat{\nu},\sigma^+}_1(\varrho) - V^{\star,\sigma^+}_1(\varrho) \right\}$.

According to the relationship in Lemma D.2, we obtain

$$\begin{aligned} V^{\star,\widehat{\nu},\sigma^+}_h(s) \leq \widehat{V}^+_h(s) &= \max_{\mu\in\Delta(\mathcal{A})} \min_{\nu\in\Delta(\mathcal{B})} \mathbb{E}_{(a,b)\sim(\mu(s),\nu(s))}\left[\widehat{Q}^+_h(s,a,b)\right] \\ &\leq \max_{\mu\in\Delta(\mathcal{A})} \mathbb{E}_{(a,b)\sim(\mu(s),\nu^\star(s))}\left[Q^+_h(s,a,b)\right], \end{aligned} \tag{50}$$

where the first equality comes from line 5 in Algorithm 2. Therefore, there exists a deterministic policy $\mu^{\mathsf{d}}: \mathcal{S} \leftarrow \Delta(\mathcal{A})$ satisfying that for any $s \in \mathcal{S}$

$$\mu^{\mathsf{d}}(s) \coloneqq \arg\max_{\mu\in\Delta(\mathcal{A})} \mathbb{E}_{(a,b)\sim(\mu(s),\nu^\star(s))}\left[Q^+_h(s,a,b)\right]. \tag{51}$$

We start by defining the following notation:

- The state-action space covered by the behavior policy $(\mu^{\mathsf{n}},\nu^{\mathsf{n}})$ in the nominal transition kernel P^0 is denoted as

$$\mathcal{C}^{\mathsf{n}} = \left\{(h,s,a,b): d^{\mathsf{n}}_h(s,a,b) > 0\right\}. \tag{52}$$

- For any time step $h \in [H]$, the sets of potential state occupancy distributions w.r.t. the policy $(\mu^{\mathsf{d}}(s), \nu^\star(s))$ and the uncertainty set $P \in \mathcal{U}^{\sigma^+}\left(P^0\right)$ are given by

$$\mathcal{D}_h^{\mathsf{p}} \coloneqq \left\{ \left[d_h^{\mu^{\mathsf{d}}(s),\nu^\star(s),P}(s) \right]_{s\in\mathcal{S}} : P \in \mathcal{U}^{\sigma^+}\left(P^0\right) \right\}; \tag{53}$$

$$\mathcal{D}_h^{\mathsf{pa}} \coloneqq \left\{ \left[d_h^{\mu^{\mathsf{d}}(s),\nu^\star(s),P}(s,a,b) \right]_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} : P \in \mathcal{U}^{\sigma^+}\left(P^0\right) \right\}. \tag{54}$$

- For convenience and without ambiguity, we introduce additional notation for $h \in [H]$ as

$$\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s) = \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \beta_h\left(s,a,b,\widehat{V}_{h+1}^+\right).$$

  In particular, the vector $\beta_h^{\mu^{\mathsf{d}},\nu^\star} \in \mathbb{R}^S$ is defined with its s-th term given by $\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s)$.
- Similarly, we can define the notation related to rewards for $h \in [H]$ as

$$\widehat{r}_h^{\mu^{\mathsf{d}},\nu^\star}(s) = \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \widehat{r}_h\left(s,a,b\right).$$

According to the update rule in line 4 in Algorithm 2 and robust Bellman equality (28), we derive

$$\begin{aligned}
&V_h^{\star,\widehat{\nu},\sigma^+}(s) - V_h^{\star,\sigma^+}(s) \qquad (55)\\
\leq& \widehat{V}_h^+(s) - V_h^{\mu^{\mathsf{d}},\nu^\star,\sigma^+}(s)\\
\leq& \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ + \beta_h^{\mu^{\mathsf{d}},\nu^\star}(s)\\
&- \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+}\\
\leq& \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \left[\inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ - \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+} \right.\\
&\left. + \left| \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ - \inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ \right| \right] + \beta_h^{\mu^{\mathsf{d}},\nu^\star}(s)\\
\overset{\text{(i)}}{\leq}& \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \left[\inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ - \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+} \right] + 2\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s)\\
\overset{\text{(ii)}}{\leq}& \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \left[P_{h,s,a,b}^{\inf,V}\left(\widehat{V}_{h+1}^+ - V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+}\right) \right] + 2\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s). \qquad (56)
\end{aligned}$$

Here, (i) holds due to (48) in Lemma D.1 for $N_h(s,a,b) > 0$ and

$$\left| \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ - \inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ \right| \leq H = \beta_h^{\mu^{\mathsf{d}},\nu^\star}(s) \text{ for } N_h(s,a,b) = 0. \tag{57}$$

Moreover, (ii) is due to

$$P_{h,s,a,b}^{\inf,V} \coloneqq \operatorname{argmin}_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+} \tag{58}$$

and consequently,

$$\inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+} = P_{h,s,a,b}^{\inf,V} V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+}, \text{ and } \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ \leq P_{h,s,a,b}^{\inf,V}\widehat{V}_{h+1}^+.$$

For ease of exposure, we introduce notation as $\widetilde{P}_{h,s}^{\inf,V} \coloneqq \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} P_{h,s,a,b}^{\inf,V}$. Furthermore, we define a sequence of matrices $\widetilde{P}_h^{\inf,V} \in \mathbb{R}^{S\times S}$. We can utilize (56) recursively over the time steps

$h, h+1, \cdots, H$ and derive

$$
\begin{aligned}
V_h^{\star,\widehat{\nu},\sigma^+}(s) - V_h^{\star,\sigma^+}(s) &\le \widehat{V}_h^+(s) - V_h^{\mu^{\mathsf{d}},\nu^\star,\sigma^+}(s) \\
&\le \widetilde{P}_h^{\inf,V}\left(\widehat{V}_{h+1}^+ - V_{h+1}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+}\right) + 2\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s) \\
&\le \widetilde{P}_h^{\inf,V}\widetilde{P}_{h+1}^{\inf,V}\left(\widehat{V}_{h+2}^+ - V_{h+2}^{\mu^{\mathsf{d}},\nu^\star,\sigma^+}\right) + 2\widetilde{P}_h^{\inf,V}\beta_{h+1}^{\mu^{\mathsf{d}},\nu^\star} + 2\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s) \\
&\le \cdots \le 2\sum_{i=h+1}^{H}\left(\prod_{j=h}^{i-1}\widetilde{P}_j^{\inf,V}\right)\beta_i^{\mu^{\mathsf{d}},\nu^\star} + 2\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s) \qquad (59) \\
&= 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widetilde{P}_j^{\inf,V}\right)\beta_i^{\mu^{\mathsf{d}},\nu^\star}, \qquad (60)
\end{aligned}
$$

For conciseness, we define $\left(\prod_{j=h}^{h-1}\widetilde{P}_j^{\inf,V}\right) = I$ to simplify notation.

For any $d_h^{\mu^{\mathsf{d}},\nu^\star} \in \mathcal{D}_h^{\mathsf{p}}$ (cf. (53)), taking inner product with (59) yields

$$
\left\langle d_h^{\mu^{\mathsf{d}},\nu^\star}, V_h^{\star,\widehat{\nu},\sigma^+} - V_h^{\star,\sigma^+}\right\rangle \le \left\langle d_h^{\mu^{\mathsf{d}},\nu^\star}, 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widetilde{P}_j^{\inf,V}\right)\beta_i^{\mu^{\mathsf{d}},\nu^\star}\right\rangle = 2\sum_{i=h}^{H}\left\langle d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}, \beta_i^{\mu^{\mathsf{d}},\nu^\star}\right\rangle, \tag{61}
$$

where

$$
d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star} \coloneqq \left[\left(d_h^{\mu^{\mathsf{d}},\nu^\star}\right)^\top\left(\prod_{j=h}^{i-1}\widetilde{P}_j^{\inf,V}\right)\right]^\top \in \mathcal{D}_i^{\mathsf{p}} \tag{62}
$$

by the definition of $\mathcal{D}_i^{\mathsf{p}}$ (cf. (53)) for all $i = h+1, \cdots, H$.

Next, we control $\langle d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}, \beta_i^{\mu^{\mathsf{d}},\nu^\star}\rangle$ by using concentrability. According to (18) in Lemma D.1, we first demonstrate that the pessimistic penalty satisfies

$$
\begin{aligned}
\beta_i(s,a,b,\hat{V}) &\le \max\left\{\sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{i,s,a,b}}(\widehat{V})}, \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_i(s,a,b)}\right\} \\
&\le \sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{i,s,a,b}}(\widehat{V})} + \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_i(s,a,b)} \\
&\overset{\text{(i)}}{\le} \sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,a,b)}\left(2\mathsf{Var}_{P^0_{i,s,a,b}}(\widehat{V}) + \frac{C_0H^2}{N_i(s,a,b)}\log\frac{KH}{\delta}\right)} + \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_i(s,a,b)} \\
&\overset{\text{(ii)}}{\le} \sqrt{\frac{2C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,a,b)}\mathsf{Var}_{P^0_{i,s,a,b}}(\widehat{V})} + \frac{\left(2C_{\mathsf{n}} + \sqrt{C_{\mathsf{n}}C_0}\right)H\log\frac{KH}{\delta}}{N_i(s,a,b)} \qquad (63)
\end{aligned}
$$

where (i) holds by applying (47) for some sufficiently large C_0 and (ii) follows from the Cauchy-Schwarz inequality. Therefore, combining the definition of $\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s)$, we obtain

$$
\begin{aligned}
\langle d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}, \beta_i^{\mu^{\mathsf{d}},\nu^\star}\rangle &= \sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s) \\
&= \sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))}\beta_i(s,a,b,\hat{V}) \\
&= \sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\,\mathbb{1}\{a=\mu^{\mathsf{d}}(s)\}\nu^\star(b|s)\beta_i(s,a,b,\hat{V}) \\
&= \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\hat{V}), \qquad (64)
\end{aligned}
$$

where the last equation holds due to the definition in (5). Then, we observe $d_h^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,a,b) \in \mathcal{D}_h^{\mathsf{pa}}$; cf. (54). Thereafter, we divide the bound (64) into two cases. *In the first case*, i.e., $s \in S$ where $\max_{P\in\mathcal{U}^{\sigma^+}(P^0)} d_i^{\mu^{\mathsf{d}},\nu^\star,P}\big(s,\mu^{\mathsf{d}}(s),b\big) = 0$, it follows from the definition (53) that: For any $d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b) \in \mathcal{D}_i^{\mathsf{pa}}$,

$$d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b) = 0. \tag{65}$$

In the second case, i.e., $s \in S$ where $\max_{P\in\mathcal{U}^{\sigma^+}(P^0)} d_i^{\mu^{\mathsf{d}},\nu^\star,P}\big(s,\mu^{\mathsf{d}}(s),b\big) > 0$, under the assumption in (20),

$$\max_{P\in\mathcal{U}^{\sigma^+}(P^0)} \frac{\min\big\{d_i^{\mu^{\mathsf{d}},\nu^\star,P}\big(s,\mu^{\mathsf{d}}(s),b\big), \frac{1}{S(A+B)}\big\}}{d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big)} \le C_{\mathsf{r}}^\star < \infty,$$

which implies that

$$d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big) > 0 \quad \text{and} \quad \big(i,s,\mu^{\mathsf{d}}(s),b\big) \in \mathcal{C}^{\mathsf{n}}. \tag{66}$$

Lemma 3.1 tells that with probability of at least $1-8\delta$,

$$\begin{aligned} N_i\big(s,\mu^{\mathsf{d}}(s),b\big) &\ge \frac{K d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big)}{8} - 5\sqrt{K d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big)\log\frac{KH}{\delta}} \\ &\overset{\text{(i)}}{\ge} \frac{K d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big)}{16} \\ &\overset{\text{(ii)}}{\ge} \frac{K\max_{P\in\mathcal{U}^{\sigma}(P^0)}\min\Big\{d_i^{\mu^{\mathsf{d}},\nu^\star,P}\big(s,\mu^{\mathsf{d}}(s),b\big), \frac{1}{S(A+B)}\Big\}}{16C_{\mathsf{r}}^\star} \\ &\ge \frac{K\min\Big\{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b), \frac{1}{S(A+B)}\Big\}}{16C_{\mathsf{r}}^\star}, \end{aligned} \tag{67}$$

Here, (i) holds since, with $f(\sigma^+,\sigma^-,H) = \min\Big\{\frac{H\sigma^+ +1-(1-\sigma^+)^H}{(\sigma^+)^2}, \frac{H\sigma^- +1-(1-\sigma^-)^H}{(\sigma^-)^2}, H\Big\}$,

$$\begin{aligned} K d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big) &\ge c_0\frac{HS(A+B)}{d_{\mathsf{m}}^{\mathsf{n}}}\log\frac{KH}{\delta} f(\sigma^+,\sigma^-,H) d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big) \\ &\ge c_0 HS(A+B)\log\frac{KH}{\delta} f(\sigma^+,\sigma^-,H) \ge 1600\log\frac{KH}{\delta}, \end{aligned} \tag{68}$$

where the first inequality follows from condition (22), and the second inequality follows from

$$d_{\mathsf{m}}^{\mathsf{n}} = \min_{h,s,\mu^{\mathsf{d}}(s),b}\big\{d_h^{\mathsf{n}}(s,\mu^{\mathsf{d}}(s),b) : d_h^{\mathsf{n}}(s,\mu^{\mathsf{d}}(s),b) > 0\big\} \le d_i^{\mathsf{n}}\big(s,\mu^{\mathsf{d}}(s),b\big). \tag{69}$$

Moreover, (ii) comes from Assumption 4.1.

Combining (63) and (64), we arrive at

$$
\begin{aligned}
&\langle d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}, \beta_i^{\mu^{\mathsf{d}},\nu^\star}\rangle \\
&= \sum_{(s,b)in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\hat{V}) \\
&\le \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\sqrt{\frac{2C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i\left(s,\mu^{\mathsf{d}}(s),b\right)}\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\left(\widehat{V}\right)} \\
&\quad + \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\frac{\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_0}\right)H\log\frac{KH}{\delta}}{N_i\left(s,\mu^{\mathsf{d}}(s),b\right)} \\
&\overset{\text{(i)}}{\le} \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\underbrace{\sqrt{\frac{32C_{\mathsf{r}}^\star C_{\mathsf{n}}\log\frac{KH}{\delta}}{K\min\left\{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b),\frac{1}{S(A+B)}\right\}}\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\left(\widehat{V}\right)}}_{B_1} \\
&\quad + \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\underbrace{\frac{16C_{\mathsf{r}}^\star\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_0}\right)H\log\frac{KH}{\delta}}{K\min\left\{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b),\frac{1}{S(A+B)}\right\}}}_{B_2}. \qquad (70)
\end{aligned}
$$

From (61), we need to bound $\sum_{i=1}^H\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_1$ and $\sum_{i=1}^H\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_2$:

D.2.1 Bounding $\sum_{i=1}^H\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_1$

Combining (68) with $\sum_{i=1}^H\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_1$ yields

$$
\begin{aligned}
&\sum_{i=1}^H\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_1 \\
=&\sum_{i=1}^H\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\sqrt{\frac{32C_{\mathsf{r}}^\star C_{\mathsf{n}}\log\frac{KH}{\delta}}{K\min\left\{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b),\frac{1}{S(A+B)}\right\}}\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\left(\widehat{V}\right)} \\
\le&\sum_{i=1}^H\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\times \\
&\max\left\{\sqrt{\frac{32C_{\mathsf{r}}^\star C_{\mathsf{n}}\log\frac{KH}{\delta}}{Kd_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)}\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\left(\widehat{V}\right)},\sqrt{\frac{32C_{\mathsf{r}}^\star C_{\mathsf{n}}S(A+B)\log\frac{KH}{\delta}}{K}\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\left(\widehat{V}\right)}\right\}, \qquad (71)
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_1\\
\leq&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}}\sqrt{\frac{32C_{\mathsf{r}}^\star C_{\mathsf{n}}\log\frac{KH}{\delta}}{K}d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\big(\widehat{V}\big)}\\
&+\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\sqrt{\frac{32C_{\mathsf{r}}^\star C_{\mathsf{n}}S(A+B)\log\frac{KH}{\delta}}{K}\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\big(\widehat{V}\big)}\\
\leq&\sqrt{\frac{32C_{\mathsf{r}}^\star C_{\mathsf{n}}S(A+B)\log\frac{KH}{\delta}}{K}}\Bigg(\sqrt{H\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\big(\widehat{V}\big)}\\
&+\sqrt{\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\big(\widehat{V}\big)}\times\sqrt{\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)}\Bigg)\\
=&\sqrt{\frac{128C_{\mathsf{r}}^\star C_{\mathsf{n}}HS(A+B)\log\frac{KH}{\delta}}{K}\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\big(\widehat{V}\big)},
\end{aligned}
\tag{72}
$$

where the last inequality follows from the Cauchy-Schwarz inequality.

Then, we introduce the following lemma about $\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P_{i,s,\mu^{\mathsf{d}}(s),b}}\big(\widehat{V}\big)$, with its proof in Appendix G.4.

Lemma D.4. *Considering $\forall\delta\in(0,1)$, with probability at least $1-\delta$, one has: for any product policy $(\widehat{\mu},\widehat{\nu})$,*

$$
\begin{aligned}
\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^0_{i,s,a,b}}\big(\widehat{V}_{i+1}\big)\leq H\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}\\
\times\left(4\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})+(H+3)\right).
\end{aligned}
\tag{73}
$$

Armed with Lemma D.4, (72) can be further bounded as

$$
\begin{aligned}
&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_1\\
&\leq\sqrt{\frac{128C_{\mathsf{r}}^\star C_{\mathsf{n}}HS(A+B)\log\frac{KH}{\delta}}{K}}\sqrt{H\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}}\\
&\quad\times\sqrt{\left(4\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})+(H+3)\right)}.
\end{aligned}
\tag{74}
$$

D.2.2 Bounding $\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_2$

Combining (67) with $\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_2$ yields

$$\begin{aligned}
&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)B_2\\
=&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\frac{16C_{\mathsf{r}}^\star\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_3}\right)H\log\frac{KH}{\delta}}{K\min\left\{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b),\frac{1}{S(A+B)}\right\}}\\
\overset{\text{(i)}}{\leq}&\frac{32C_{\mathsf{r}}^\star\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_3}\right)H^2S(A+B)\log\frac{KH}{\delta}}{K},
\end{aligned}\tag{75}$$

where the inequality holds by the trivial fact

$$\begin{aligned}
&\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}}\frac{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)}{\min\left\{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b),\frac{1}{S(A+B)}\right\}}\\
\leq&\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\left(\frac{1}{d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)}+\frac{1}{1/S(A+B)}\right)\\
=&\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} 1+S(A+B)\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\leq 2S(A+B).
\end{aligned}\tag{76}$$

D.2.3 Putting all together

Combining (74) and (75), we obtain

$$\begin{aligned}
&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})\\
\leq&\sqrt{\frac{128C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}}\\
&\times\sqrt{\left(4\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})+(H+3)\right)}\\
&+\frac{32C_{\mathsf{r}}^\star\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_3}\right)H^2S(A+B)\log\frac{KH}{\delta}}{K},
\end{aligned}\tag{77}$$

which can further bound as

$$
\begin{aligned}
&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})\\
\leq&\sqrt{\frac{128C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2(H+3)S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}}\\
&+\frac{32C_{\mathsf{r}}^\star\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_3}\right)H^2S(A+B)\log\frac{KH}{\delta}}{K}+\sqrt{\frac{512C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2S(A+B)\log\frac{KH}{\delta}}{K}}\\
&\times\sqrt{\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})}\\
\leq&\sqrt{\frac{128C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2(H+3)S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}}\\
&+\frac{32C_{\mathsf{r}}^\star\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_3}\right)H^2S(A+B)\log\frac{KH}{\delta}}{K}\\
&+\frac{256C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}\\
&+\frac{1}{2}\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V}),
\end{aligned}
\tag{78}
$$

where the last relation follows from the AM-GM inequality. Rearranging the terms, it follows that

$$
\begin{aligned}
&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})\\
\leq&\sqrt{\frac{512C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2(H+3)S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}}\\
&+\frac{64C_{\mathsf{r}}^\star\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_3}\right)H^2S(A+B)\log\frac{KH}{\delta}}{K}\\
&+\frac{512C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}\\
\leq&\sqrt{\frac{512C_{\mathsf{r}}^\star C_{\mathsf{n}}H^2(H+3)S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}}\\
&+\frac{C_{\mathsf{r}}^\star C_2H^2S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}.
\end{aligned}
\tag{79}
$$

Along with the above analysis, we are ready to bound $V_1^{\star,\sigma^+}(\varrho)-V_1^{\widehat{\mu},\star,\sigma^+}(\varrho)$: There exist some sufficiently large constants $C_1,C_2,C_3>0$, and

$$
\begin{aligned}
V_1^{\star,\widehat{\nu},\sigma^+}(\varrho)-V_1^{\star,\sigma^+}(\varrho)\leq&\sqrt{\frac{C_{\mathsf{r}}^\star C_1H^3S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}}\\
&+\frac{C_{\mathsf{r}}^\star C_2H^2S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}\\
\leq&\sqrt{\frac{C_{\mathsf{r}}^\star C_3H^3S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}},
\end{aligned}
\tag{80}
$$

where the last inequality follows from condition (22).

D.3 Step 3: Summing up

Consequently, we obtain the upper bound of $V_1^{\star,\widehat{\nu},\sigma^+}(\varrho) - V_1^{\widehat{\mu},\widehat{\nu},\sigma^+}(\varrho)$ in (80). Similarly,

$$\begin{aligned} &V_1^{\star,\sigma^-}(\varrho) - V_1^{\widehat{\mu},\star,\sigma^-}(\varrho) \\ &\leq \sqrt{\frac{C_{\mathsf{r}}^\star C_3 H^2 S(A+B)\log\frac{KH}{\delta}}{K}\min\left\{\frac{(H+1)(H\sigma^- - 1 + (1-\sigma^-)^H)}{(\sigma^-)^2}, H\right\}}, \end{aligned} \tag{81}$$

which directly leads to

$$\begin{aligned} \mathrm{Gap}(\widehat{\mu},\widehat{\nu}) &\leq c_1\sqrt{\frac{C_{\mathsf{r}}^\star H^2 S(A+B)\log\frac{KH}{\delta}}{K}} \\ &\quad\times\sqrt{\min\left\{\frac{2(H\sigma^+ - 1 + (1-\sigma^+)^H)}{(\sigma^+)^2}, \frac{2(H\sigma^- - 1 + (1-\sigma^-)^H)}{(\sigma^-)^2}, H\right\}}, \end{aligned} \tag{82}$$

for some sufficiently large c_1 and

$$K \geq HS(A+B)\log\frac{KH}{\delta}\min\left\{\frac{2(H\sigma^+ - 1 + (1-\sigma^+)^H)}{(\sigma^+)^2}, \frac{2(H\sigma^- - 1 + (1-\sigma^-)^H)}{(\sigma^-)^2}, H\right\}.$$

Discussion of (82). For the term $T = \min(f(\sigma^+,\sigma^-), H)$, considering the interchangeability between σ^+ and σ^-, we define $g(\sigma^+, H) = H\sigma^+ - H(1-\sigma^+)^H - (\sigma^+)^2 H$. For $H \geq 2$, we derive the first derivative as $\frac{\partial g(\sigma^+,H)}{\partial \sigma^+} = H + H^2(1-\sigma^+)^{H-1} - 2H\sigma^+$. The second derivative is given by $\frac{\partial^2 g(\sigma^+,H)}{\partial(\sigma^+)^2} = -H^2(H-1)(1-\sigma^+)^{H-2} - 2H < 0$, indicating that $g(\sigma^+, H)$ is concave. By evaluating the first derivative at the boundaries, we find $\frac{\partial g(\sigma^+,H)}{\partial \sigma^+}|_{\sigma^+\to 0} \to H^2 + H > 0$ and $\frac{\partial g(\sigma^+,H)}{\partial \sigma^+}|_{\sigma^+=1} = -H < 0$, showing that $g(\sigma^+, H)$ first increases monotonically, reaches a maximum at some point $\sigma^\star$, and then decreases monotonically. Furthermore, since $g(\sigma^+ \to 0, H) \to -H < 0$ and $g(\sigma^+ = 1, H) = 0$, there exists $0 < \sigma^0 < 1$ such that $g(\sigma^0, H) = 0$. If $\sigma^0 \lesssim \min\{\sigma^+,\sigma^-\} \lesssim 1$, we have $T = H$. Otherwise, $T = \min\left\{\frac{(H\sigma^+ - 1 + (1-\sigma^+)^H)}{(\sigma^+)^2}, \frac{(H\sigma^- - 1 + (1-\sigma^-)^H)}{(\sigma^-)^2}\right\}$.

E Proof of Theorem 4.4

We focus on a simple class of RTZMGs: robust Markov decision processes (RMDPs), which are single-agent versions of RTZMGs. Recall that an RTZMG with an uncertainty set is $\mathcal{MG} = \{\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{U}^{\sigma^+}(P^0), \mathcal{U}^{\sigma^-}(P^0), r, H\}$. For illustration convenience, we assume $\mathcal{A} \geq \mathcal{B}$ and $|\mathcal{B}| = 1$, meaning the min-player's actions do not affect transitions or rewards. Thus, finding a robust NE in RTZMGs reduces to finding the max-player's optimal policy in a corresponding RMDP $\mathcal{M}_{\mathsf{r}} = \{\mathcal{S}, \mathcal{A}, \mathcal{U}^{\sigma^+}(P^0), r, H\}$.

Thus, we construct the lower bound for finding the optimal policy in RTZMGs, which also implies a lower bound for finding robust NE in RTZMGs. We start by re-stating a useful property about KL divergence from Lemma 2.7 in [36].

Lemma E.1. *$\forall p, q \in (0,1)$, it holds that*

$$\mathsf{KL}(p \,\|\, q) \leq \frac{(p-q)^2}{q(1-q)}. \tag{83}$$

E.1 Step 1: Constructing a family of hard Markov game instances

The hard instances developed here differ from standard MDP since we need to consider that the transition kernel can be perturbed in robust MDPs.

E.1.1 Constructing hard robust MDP instances

We introduce an auxiliary collection $\Phi \subseteq \{0,1\}^H$ comprising H-dimensional vectors. Resorting to the Gilbert-Varshamov lemma [15], there also exists a set $\Phi \subseteq \{0,1\}^H$ such that:

$$\text{for any } \phi, \widetilde{\phi} \in \Phi \text{ obeying } \phi \neq \widetilde{\phi}: \quad \|\phi - \widetilde{\phi}\|_1 \geq \frac{H}{8} \qquad \text{and} \qquad |\Phi| \geq e^{H/8}. \tag{84}$$

Bearing this in mind, we construct a set of RMDPs as

$$\mathcal{M}(\mathcal{F}, \Phi) \coloneqq \left\{ \mathcal{M}_f^\phi = \big(\mathcal{S}, \mathcal{A}, \mathcal{U}^{\sigma^+}(P^{f,\phi}), r, H\big) \,|\, f \in \mathcal{F} = \{0, 1, \cdots, SA-1\}, \phi = [\phi_h]_{1 \leq h \leq H} \in \Phi \right\}, \tag{85}$$

where $\mathcal{S} = \{0, 1, \ldots, S-1\}$, $\mathcal{A} = \{0, 1, \cdots, A-1\}$, and σ^+ will be introduced momentarily.

In simple terms, the collection $\mathcal{M}(\mathcal{F}, \Phi)$ consists of SA subsets, each containing $|\Phi|$ different RMDPs associated with some $f \in \mathcal{F}$. The state space for each RMDP $\mathcal{M}_f^\phi \in \mathcal{M}(\mathcal{F}, \Phi)$, denoted as $\mathcal{S}_{\mathsf{one}}$, includes two types of states: $\mathcal{M} = \{m_i \mid i \in \mathcal{F}\}$ and $\mathcal{N} = \{n_i \mid i \in \mathcal{F}\}$. Each state in $\mathcal{M}$ and $\mathcal{N}$ has two possible actions, $\mathcal{A}_{\mathsf{one}} = \{0, 1\}$. Thus, there are a total of $2SA$ states and $4SA$ state-action pairs.

Now, we can define the transition kernels for $\mathcal{M}(\mathcal{F}, \Phi)$. For any RMDP $\mathcal{M}_f^\phi \in \mathcal{M}(\mathcal{F}, \Phi)$, the transition kernel $P^{f,\phi} = \{P_h^{f,\phi}\}_{h=1}^H$ is defined as follows, for any $(s, a, s', h) \in \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}} \times \mathcal{S}_{\mathsf{one}} \times [H]$,

$$P_h^{f,\phi}(s' \,|\, s, a) = \begin{cases} p\mathbb{1}(s' = n_f) + (1-p)\mathbb{1}(s' = s), & \text{if} \quad s = m_f, a = \phi_h; \\ q\mathbb{1}(s' = n_f) + (1-q)\mathbb{1}(s' = s), & \text{if} \quad s = m_f, a = 1 - \phi_h; \\ \mathbb{1}(s' = s), & \text{otherwise}, \end{cases} \tag{86}$$

where $p > q \geq \frac{1}{2}$.

In addition, the reward function is defined as

$$\forall (h, s, a) \in [H] \times \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}}: \quad r_h(s, a) = \begin{cases} 1, & \text{if } s \in \mathcal{N}; \\ 0, & \text{otherwise.} \end{cases} \tag{87}$$

E.1.2 Uncertainty set of the transition kernels

Denote the transition kernel vector as

$$\forall (h, s, a) \in [H] \times \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}}: \quad P_{h,s,a}^{f,\phi} \coloneqq P_h^{f,\phi}(\cdot \,|\, s, a) \in \Delta(\mathcal{S}). \tag{88}$$

Recall the uncertainty set defined in (1). $\mathcal{U}^{\sigma^+}(P^{f,\phi})$ represents

$$\mathcal{U}^{\sigma^+}(P^{f,\phi}) \coloneqq \otimes \mathcal{U}^{\sigma^+}(P_{h,s,a}^{f,\phi}), \quad \mathcal{U}^{\sigma^+}(P_{h,s,a}^{f,\phi}) \coloneqq \left\{ \widetilde{P}_{h,s,a}^{f,\phi} \in \Delta(\mathcal{S}) : \rho\left(\widetilde{P}_{h,s,a}^{f,\phi} - P_{h,s,a}^{f,\phi}\right) \leq \sigma^+ \right\},$$

where $\otimes$ represents the Cartesian product over $(h, s, a) \in [H] \times \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}}$. For the convenience of the subsequent proof, we analyze the TV distance as an uncertainty set for example, which means

$$\mathcal{U}^{\sigma^+}(P_{h,s,a}^{f,\phi}) \coloneqq \left\{ \widetilde{P}_{h,s,a}^{f,\phi} \in \Delta(\mathcal{S}) : \frac{1}{2} \left\| \widetilde{P}_{h,s,a}^{f,\phi} - P_{h,s,a}^{f,\phi} \right\| \leq \sigma^+ \right\}. \tag{89}$$

For any RMDP $\mathcal{M}_f^\phi \in \mathcal{M}(\mathcal{F}, \Phi)$ and any $(h, s, a, s') \in [H] \times \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}} \times \mathcal{S}_{\mathsf{one}}$, we define the minimum transition probability from (s, a) to s', determined by any perturbed transition kernel $P_{h,s,a} \in \mathcal{U}^{\sigma^+}(P_{h,s,a}^{f,\phi})$, as

$$P_h^{\inf, f, \phi}(s' \,|\, s, a) \coloneqq \inf_{P_{h,s,a} \in \mathcal{U}^{\sigma^+}(P_{h,s,a}^{f,\phi})} P_h(s' \,|\, s, a) = \max\{P_h(s' \,|\, s, a) - \sigma^+, 0\}, \tag{90}$$

where the last equation inherits from the definition of $\mathcal{U}^{\sigma^+}(\cdot)$ in (89), with the remaining probability distributed to other states. We also define the transition from each $s \in \mathcal{M}$ to the corresponding state $s^{m \to n} \in \mathcal{N}$ for any $\mathcal{M}_f^\phi$: for all $h \in [H]$,

$$\begin{aligned} \text{for } m_f: \quad & p_h^{\inf} \coloneqq P_h^{\inf, f, \phi}(n_f \,|\, m_f, \phi_h) = p - \sigma^+, \\ & q_h^{\inf} \coloneqq P_h^{\inf, f, \phi}(n_f \,|\, m_f, 1 - \phi_h) = q - \sigma^+. \end{aligned} \tag{91}$$

It is obvious that

$$p_1^{\inf} = p_2^{\inf} = \cdots p_H^{\inf}, \quad q_1^{\inf} = q_2^{\inf} = \cdots q_H^{\inf}, \tag{92}$$

which motivates us to abbreviate them consistently as $p^{\inf} \coloneqq p_1^{\inf}$ and $q^{\inf} \coloneqq q_1^{\inf}$ later.

E.1.3 Robust value functions and optimal policies

We now define the robust value functions and identify the optimal policies for RMDP instances. For any RMDP $\mathcal{M}_f^\phi \in \mathcal{M}(\mathcal{F}, \Phi)$, let $\widetilde{\mu}^{\star,f,\phi} = \{\mu_h^{\star,f,\phi}\}_{h=1}^H$ represent the optimal policy, given that ν is deterministic. At each step h, we use $V_h^{\widetilde{\mu},\sigma^+,f,\phi}$ and $V_h^{\star,\sigma^+,f,\phi}$ to denote the robust value function of any policy $\widetilde{\mu}$ and the optimal policy $\widetilde{\mu}^{\star,f,\phi}$, respectively, under uncertainty level σ^+. The following lemma highlights key properties of robust value functions and optimal policies; the proof is deferred to Appendix H.1.

Lemma E.2. *Consider any $\mathcal{M}_f^\phi \in \mathcal{M}(\mathcal{F}, \Phi)$ and any policy $\widetilde{\mu}$. Defining*

$$m_h^{\widetilde{\mu},f,\phi} = p^{\inf}\widetilde{\mu}_h(\phi_h \,|\, m_f) + q^{\inf}\widetilde{\mu}_h(1-\phi_h \,|\, m_f), \tag{93}$$

it holds that

$$\forall h \in [H]: \quad V_h^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) = m_h^{\widetilde{\mu},f,\phi} V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + (1 - m_h^{\widetilde{\mu},f,\phi}) V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f), \tag{94a}$$

$$\forall (s,h) \in \mathcal{N} \times [H]: \quad V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s) = 1 + (1-\sigma^+) V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s) + \sigma^+ V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f). \tag{94b}$$

In addition, for all $h \in [H]$, the optimal policy and the optimal value function obey

$$\widetilde{\mu}_h^{\star,f,\phi}(\phi_h \,|\, m_f) = \widetilde{\mu}_h^{\star,f,\phi}(\phi_h \,|\, n_f) = 1, \tag{95}$$

$$V_h^{\star,\sigma^+,f,\phi}(m_f) = p^{\inf} V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + (1-p^{\inf}) V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f). \tag{96}$$

E.1.4 Construction of the history/batch dataset

In the nominal environment $\mathcal{M}_f^{\phi,\mathsf{n}}$, a batch dataset is generated with K independent sample trajectories with length H per trajectory, according to (3) and based on the initial state distribution ϱ^{n} and behavior policy $\widetilde{\mu}^{\mathsf{n}} = \{\mu_h^{\mathsf{n}}\}_{h=1}^H$ satisfying

$$\varrho^{\mathsf{n}}(s) = \varrho(s) \quad \text{and} \quad \widetilde{\mu}_h^{\mathsf{n}}(a \,|\, s) = \frac{1}{2}, \qquad \forall (s,a,h) \in \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}} \times [H]. \tag{97}$$

We define the nominal transition kernels for $\mathcal{M}_f^{\phi,\mathsf{n}}$, where any state $m_i \in \mathcal{M}$ transitions only to the corresponding $n_i \in \mathcal{N}$ or remains at itself. For simplicity, for any $s = m_i \in \mathcal{M}$, we denote the corresponding state $n_i \in \mathcal{N}$ as $s^{m\to n}$. The basic nominal transition kernel is defined as: For all $(h,s,a) \in [H] \times \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}}$,

$$P_h^\star(s' \,|\, s,a) = \begin{cases} (p+\Delta)\mathbb{1}(s' = s^{m\to n}) + (1-p-\Delta)\mathbb{1}(s'=s), & \text{if} \quad s \in \mathcal{M}, a = \phi_h; \\ p\mathbb{1}(s' = s^{m\to n}) + (1-p)\mathbb{1}(s'=s), & \text{if} \quad s \in \mathcal{M}, a = 1-\phi_h; \\ \mathbb{1}(s'=s), & \text{if} \quad s \in \mathcal{N}. \end{cases} \tag{98}$$

In other words, the transition kernel of each $\mathcal{M}_f^\phi \in \mathcal{M}(\mathcal{F}, \Phi)$ differs slightly from the basic nominal transition kernel $\mathcal{M}_f^{\phi,\mathsf{n}}$ when $s = m_f$, making all components within $\mathcal{M}(\mathcal{F}, \Phi)$ close to each other. Specifically, p and q are set according to

$$0 \le p \le p + \Delta \le 1 \text{ and } 0 \le q = p - \Delta \text{ for some } p, \Delta > 0. \tag{99}$$

Without loss of generality, let $\sigma^+ \in (0, 1-c_0]$ for some $0 < c_0 < 1$ be the uncertainty level. Taking $c_2 \le \frac{1}{4}$ and $c_1 \coloneqq \frac{c_0}{2} \le \frac{1}{4}$, p and Δ are set as

$$p = \begin{cases} \frac{c_2}{H}, & \text{if } \sigma^+ \le \frac{c_2}{H} \\ \left(1 + \frac{c_1}{H}\right)\sigma^+ & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Delta \le \begin{cases} \frac{c_2}{2H}, & \text{if } \sigma^+ \le \frac{c_2}{2H} \\ \frac{c_1}{H}\sigma^+ & \text{otherwise} \end{cases} \tag{100}$$

which establishes the fact that

$$p + \Delta \geq p \geq q = p - \Delta \geq \max\left\{\frac{c_2}{2H}, \sigma^+\right\}. \tag{101}$$

Combined with $H \geq 2$, it is easily verified that $0 \leq p + \Delta \leq 1$ as follows:

$$\begin{aligned} &\text{when } \sigma^+ > \frac{c_2}{2H}: \quad \left(1 + \frac{c_1}{H}\right)\sigma^+ + \frac{c_1}{H}\sigma^+ \leq 1 - c_0 + \frac{2c_1}{H}\sigma^+ \leq 1 - \frac{c_0(H-1)}{H} < 1, \\ &\text{when } \sigma^+ \leq \frac{c_2}{2H}: \quad \frac{3c_2}{2H} \leq 1. \end{aligned} \tag{102}$$

In addition, let $\overline{\varrho}(s)$ represent a state distribution supported on the state subset $(m_f, n_f) \in \mathcal{M} \times \mathcal{N}$:

$$\overline{\varrho}(s) = \frac{1}{CSA}\mathbb{1}(s = m_f) + \Big(1 - \frac{1}{CSA}\Big)\mathbb{1}(s = n_f), \tag{103}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $C > 0$ is some constant that determines the concentrability coefficient $C_{\mathrm{r}}^\star$ (as we shall detail momentarily) and obeys

$$\frac{1}{CSA} \leq \frac{1}{4}. \tag{104}$$

As it turns out, for any MDP $\mathcal{M}_\phi^f$, the occupancy distributions of the above batch dataset are the same (due to interchangeability) and admit the following simple characterization:

$$\forall (s,a) \in \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}}, \qquad d_1^{\mathsf{n}, P^{\phi,f}}(s,a) = \frac{1}{2}\overline{\varrho}(s), \tag{105a}$$

$$\forall (s,a,h) \in \mathcal{S}_{\mathsf{one}} \times \mathcal{A}_{\mathsf{one}} \times [H], \qquad \frac{\overline{\varrho}(s)}{2} \leq d_h^{\mathsf{n}, P^{\phi,f}}(s) \leq 2\overline{\varrho}(s), \quad \frac{\overline{\varrho}(s)}{4} \leq d_h^{\mathsf{n}, P^{\phi,f}}(s,a) \leq \overline{\varrho}(s). \tag{105b}$$

In addition, we choose the following initial state distribution

$$\varrho(s) = \begin{cases} \frac{1}{CSA}, & \text{if } s \in \mathcal{M} \\ 0, & \text{if } s \in \mathcal{N}. \end{cases} \tag{106}$$

With this choice of ϱ, the single-policy clipped concentrability coefficient $C_{\mathrm{r}}^\star$ and the quantity C are intimately connected:

$$C \leq C_{\mathrm{r}}^\star \leq 2C. \tag{107}$$

The proofs of (105) and (107) are postponed to Appendix H.2 and H.3, respectively.

E.2 Step 2: Establishing the minimax lower bound

Recall our goal: for any policy estimator $\widetilde{\mu}$ computed based on the empirical dataset, we plan to control the quantity

$$\max_{(f,\phi) \in \mathcal{F} \times \Phi} \left\{ V_1^{\star, \sigma^+, f, \phi}(\varrho) - V_1^{\widetilde{\mu}, \sigma^+, f, \phi}(\varrho) \right\} \tag{108}$$

with initial state distribution defined in (106).

E.2.1 Step 1: Converting the goal to estimate (f, ϕ)

As verified in Appendix H.4, we have

$$\varepsilon \leq \begin{cases} \frac{c_2}{H}, & \text{if } \sigma^+ \leq \frac{c_2}{2H}; \\ 1, & \text{otherwise,} \end{cases} \tag{109}$$

and

$$\Delta = c_5 \begin{cases} \frac{\varepsilon}{H^2}, & \text{if } \sigma^+ \leq \frac{c_2}{2H}; \\ \frac{\sigma^+ \varepsilon}{H}, & \text{otherwise,} \end{cases} \tag{110}$$

which satisfies (100) and leads to that for any policy $\widetilde{\mu}$ obeying

$$\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1\geq\frac{H}{8}, \tag{111}$$

one has

$$V_1^{\star,\sigma^+,f,\phi}(m_f)-V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)>\varepsilon, \tag{112}$$

whose proof is postponed to Appendix H.4. Now, we are ready to convert the estimation of an optimal policy to estimating (f,ϕ). Let $\mathbb{P}_{f,\phi}$ represent the probability distribution when the RMDP is $\mathcal{M}_f^{\phi}$, $\forall(f,\phi)\in\mathcal{F}\times\Phi$. For any $(f,\phi)\in\mathcal{F}\times\Phi$, suppose that there exists a policy $\widetilde{\mu}$ achieving

$$\mathbb{P}_{f,\phi}\left\{V_1^{\star,\sigma^+,f,\phi}(m_f)-V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\leq\varepsilon\right\}\geq\frac{3}{4}, \tag{113}$$

which, in view of (112), indicates that

$$\mathbb{P}_{f,\phi}\left\{\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1<\frac{H}{8}\right\}\geq\frac{3}{4}. \tag{114}$$

Consequently, taking $\widetilde{\phi}=\arg\min_{\phi\in\Phi}\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1$, we construct the estimate of ϕ as $\widehat{\phi}=\widetilde{\phi}$. If $\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1<\frac{H}{8}$ for some $\phi\in\Phi$, then for any $\phi'\in\Phi$ obeying $\phi'\neq\phi$, one has

$$\begin{aligned}
&\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi'}(\cdot\,|\,m_f)\right\|_1\\
\geq&\sum_{h=1}^{H}\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi'}(\cdot\,|\,m_f)\right\|_1-\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1\\
>&\frac{H}{4}-\frac{H}{8}=\frac{H}{8},
\end{aligned} \tag{115}$$

where the first inequality holds due to the triangle inequality, and the last inequality follows from the assumption $\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1<\frac{H}{8}$ and the separation property of $\phi\in\Phi$; see (84). Similarly, we have $\widehat{\phi}=\phi$ if

$$\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1<\frac{H}{8}<\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi'}(\cdot\,|\,m_f)\right\|_1,\forall\phi'\in\Phi,\phi'\neq\phi, \tag{116}$$

which can be directly achieved when $\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1<\frac{H}{8}$, and further leads to

$$\mathbb{P}_{f,\phi}\left[\widehat{\phi}=\phi\right]\geq\mathbb{P}_{f,\phi}\left\{\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f)-\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1<\frac{H}{8}\right\}\geq\frac{3}{4}. \tag{117}$$

E.2.2 Step 2: Developing the probability of error in testing multiple hypotheses

Next, we address the hypothesis testing problem over $\phi\in\Phi$ and derive the information-theoretic lower bound for the probability of error. Specifically, we define the minimax probability of error as:

$$p_{\mathrm{e}}:=\inf_{(\widehat{f},\widehat{\phi})}\max_{(f,\phi)\in\mathcal{F}\times\Phi}\mathbb{P}_{f,\phi}\big(\widehat{\phi}\neq\phi\big),$$

where the infimum is taken over all possible tests $\widehat{\phi}$ constructed from the available batch dataset.

Given the dataset $\mathcal{D}_0$ with K independent trajectories, let $\varrho^{\mathsf{n},\phi}$ (and $\varrho_h^{\mathsf{n},\phi}(s,a)$) represent the distribution vector (and distribution) of each sample tuple (s_h,a_h,s_h') at time step h under the nominal

transition kernel $P^\star$ for $\mathcal{M}_f^{\phi,\mathsf{n}}$. Then, employing Fano's inequality [36, Theorem 2.2] and the additivity of KL divergence [36, Page 85], it follows that

$$\begin{aligned} p_{\mathrm{e}} &\geq 1 - K\frac{\max_{(\phi,\widetilde{\phi})\in\Phi,\phi\neq\widetilde{\phi}}\mathsf{KL}\big(\varrho^{\mathsf{n},\phi}\,|\,\varrho^{\mathsf{n},\widetilde{\phi}}\big)+\log 2}{\log|\Phi|} \\ &\overset{\text{(i)}}{\geq} 1 - \frac{8K}{H}\max_{(\phi,\widetilde{\phi})\in\Phi,\phi\neq\widetilde{\phi}}\mathsf{KL}\big(\varrho^{\mathsf{n},\phi}\,|\,\varrho^{\mathsf{n},\widetilde{\phi}}\big) - \frac{8\log 2}{H} \\ &\overset{\text{(ii)}}{\geq} \frac{1}{2} - \frac{8K}{H}\max_{(\phi,\widetilde{\phi})\in\Phi,\phi\neq\widetilde{\phi}}\mathsf{KL}\big(\varrho^{\mathsf{n},\phi}\,|\,\varrho^{\mathsf{n},\widetilde{\phi}}\big), \end{aligned} \tag{118}$$

where (i) holds due to $|\Phi| \geq e^{H/8}$ and (ii) follows from $H \geq 16\log 2$.

Since the occupancy state distribution d_h^{n} is the same for any MDP $\mathcal{M}_f^{\phi}$, $\forall\phi\in\Phi$, we apply the chain rule of KL divergence [12, Lemma 5.2.8] and the Markov property of the independent sample trajectories to obtain:

$$\begin{aligned} \mathsf{KL}\big(\varrho^{\mathsf{n},\phi}\,|\,\varrho^{\mathsf{n},\widetilde{\phi}}\big) &= \sum_{h=1}^{H}\mathop{\mathbb{E}}_{s\sim d_h^{\mathsf{n}}(s)}\Big[\mathsf{KL}\big(P_h^{\star,\phi}(\cdot\,|\,s,a)\parallel P_h^{\star,\widetilde{\phi}}(\cdot\,|\,s,a)\big)\Big] \\ &\overset{\text{(i)}}{=} \frac{1}{2}\overline{\varrho}(m_f)\sum_{h=1}^{H}\sum_{a\in\{0,1\}}\Big[\mathsf{KL}\big(P_h^{\phi}(\cdot\,|\,m_f,a)\parallel P_h^{\widetilde{\phi}}(\cdot\,|\,m_f,a)\big)\Big], \end{aligned} \tag{119}$$

where (i) follows from applying (105) and obtaining

$$\begin{aligned} &\mathop{\mathbb{E}}_{s\sim d_h^{\mathsf{n}}(s)}\Big[\mathsf{KL}\big(P_h^{\star,\phi}(\cdot\,|\,s,a)\parallel P_h^{\star,\widetilde{\phi}}(\cdot\,|\,s,a)\big)\Big] \\ =&\sum_s d_h^{\mathsf{n}}(s)\left\{\sum_{a,s'}\widetilde{\mu}_h^{\mathsf{n}}(a\,|\,s)P_h^{\phi_h}(s'\,|\,s,a)\log\frac{\widetilde{\mu}_h^{\mathsf{n}}(a\,|\,s)P_h^{\phi_h}(s'\,|\,s,a)}{\widetilde{\mu}_h^{\mathsf{n}}(a\,|\,s)P_h^{\widetilde{\phi}_h}(s'\,|\,s,a)}\right\} \\ =&\frac{1}{2}\overline{\varrho}(m_f)\sum_a\sum_{s'}P_h^{\phi_h}(s'\,|\,m_f,a)\log\frac{P_h^{\phi_h}(s'\,|\,m_f,a)}{P_h^{\widetilde{\phi}_h}(s'\,|\,m_f,a)} \\ =&\frac{1}{2}\overline{\varrho}(m_f)\sum_a\mathsf{KL}\big(P_h^{\phi_h}(\cdot\,|\,m_f,a)\parallel P_h^{\widetilde{\phi}_h}(\cdot\,|\,m_f,a)\big). \end{aligned}$$

Consequently, combining (118) and (119) leads to

$$p_{\mathrm{e}} \geq \frac{1}{2} - \frac{4K}{H}\max_{(\phi,\widetilde{\phi})\in\Phi,\phi\neq\widetilde{\phi}}\left[\overline{\varrho}(m_f)\sum_{h=1}^{H}\sum_a\mathsf{KL}\big(P_h^{\phi_h}(\cdot\,|\,m_f,a)\parallel P_h^{\widetilde{\phi}_h}(\cdot\,|\,m_f,a)\big)\right]. \tag{120}$$

We proceed to analyze (120) by considering different cases of the uncertainty level σ^+.

- For $0<\sigma^+\leq\frac{c_2}{2H}$: If $\phi_h=\widetilde{\phi}_h$, it is obvious that

$$\sum_{a\in\{0,1\}}\mathsf{KL}\big(P_h^{\star,\phi}(\cdot\,|\,s,a)\parallel P_h^{\star,\widetilde{\phi}}(\cdot\,|\,s,a)\big)=0. \tag{121}$$

  Consider the case of $\phi_h\neq\widetilde{\phi}_h$. Without loss of generality, we suppose $\phi_h=0$ and $\widetilde{\phi}_h=1$, which indicates

$$\begin{aligned} \mathsf{KL}\big(P_h^{\star,\phi}(0\,|\,m_f,0)\parallel P_h^{\star,\widetilde{\phi}}(0\,|\,m_f,0)\big) &\leq \frac{(p-q)^2}{q(1-q)} \overset{\text{(i)}}{=} \frac{\Delta^2}{q(1-q)} \\ &\overset{\text{(ii)}}{=} \frac{(c_5)^2\varepsilon^2}{H^4q(1-q)} \leq \frac{4(c_5)^2\varepsilon^2}{c_2H^3}, \end{aligned} \tag{122}$$

  where the first inequality exists by applying Lemma E.1, (i) follows from the definitions in (99), (ii) holds due to the definition in (110), and the last inequality arises from $q=p-\Delta\geq$

$\frac{c_2}{2H}$ (see (100)) and $1-q \geq 1-p \geq 1-\frac{c_2}{H} \geq \frac{1}{2}$. Similarly, we establish the same bound for $\mathsf{KL}\big(P_h^{\star,\phi}(0\,|\,m_f,1) \,\|\, P_h^{\star,\widetilde{\phi}}(0\,|\,m_f,1)\big)$. Further incorporating (122) yields

$$\sum_{a\in\{0,1\}} \mathsf{KL}\big(P_h^{\star,\phi}(\cdot\,|\,m_f,a) \,\|\, P_h^{\star,\widetilde{\phi}}(\cdot\,|\,m_f,a)\big) \leq \frac{16(c_5)^2\varepsilon^2}{c_2H^3}. \tag{123}$$

- For $\frac{c_2}{2H} < \sigma^+ \leq 1-c_0$: Following the same pipeline, it then boils down to control the main term as below:

$$\begin{aligned}\mathsf{KL}\big(P_h^{\star,\phi}(0\,|\,m_f,0) \,\|\, P_h^{\star,\widetilde{\phi}}(0\,|\,m_f,0)\big) &\leq \frac{(p-q)^2}{q(1-q)} \overset{\text{(i)}}{=} \frac{\Delta^2}{q(1-q)} \\ &\overset{\text{(ii)}}{=} \frac{(c_5)^2{\sigma^+}^2\varepsilon^2}{H^2q(1-q)} \leq \frac{2(c_5)^2\sigma^+\varepsilon^2}{c_0H^2},\end{aligned} \tag{124}$$

where (i) and (ii) follow from the definitions in (99) and (110). Here, the last inequality arises from

$$\begin{aligned}1-q \geq 1-p = 1-(1+\frac{c_1}{H})\sigma^+ &\overset{\text{(i)}}{\geq} c_0 - \frac{c_1}{H} \overset{\text{(ii)}}{\geq} \frac{c_0}{2} \\ p \geq q = p-\Delta &\overset{\text{(iii)}}{\geq} \sigma^+,\end{aligned} \tag{125}$$

where (ii) holds due to the definition of $c_1 = \frac{c_0}{2}$, and (iii) follows from (101). Consequently, we arrive at

$$\sum_{a\in\{0,1\}} \mathsf{KL}\big(P_h^{\star,\phi}(\cdot\,|\,,s,a) \,\|\, P_h^{\star,\widetilde{\phi}}(\cdot\,|\,,s,a)\big) \leq \frac{8(c_5)^2\sigma^+\varepsilon^2}{c_0H^2}. \tag{126}$$

Summing up (123) and (126), we achieve for any $(\phi,\widetilde{\phi}) \in \Phi$ with $\phi \neq \widetilde{\phi}$ and any time step $h \in [H]$

$$\sum_{a\in\{0,1\}} \mathsf{KL}\big(P_h^{\star,\phi}(\cdot\,|\,m_f,a) \,\|\, P_h^{\star,\widetilde{\phi}}(\cdot\,|\,m_f,a)\big) \leq \frac{16(c_5)^2\varepsilon^2}{c_0c_2H^2}\max\{\sigma^+,1/H\}. \tag{127}$$

Plugging (127) back to (120), under the definition in (106), we obtain

$$\begin{aligned}p_{\mathrm{e}} &\geq \frac{1}{2} - \frac{4K}{H}\max_{(\phi,\widetilde{\phi})\in\Phi,\phi\neq\widetilde{\phi}}\left[\overline{\varrho}(m_f)\sum_{h=1}^{H}\sum_{a}\mathsf{KL}\big(P_h^{\phi_h}(\cdot\,|\,m_f,a) \,\|\, P_h^{\widetilde{\phi}_h}(\cdot\,|\,m_f,a)\big)\right] \\ &\geq \frac{1}{2} - \frac{4K}{H}\overline{\varrho}(m_f)\sum_{h=1}^{H}\frac{16(c_5)^2\varepsilon^2}{c_0c_2H^2}\max\{\sigma^+,1/H\} \\ &\geq \frac{1}{2} - \frac{64K(c_5)^2\varepsilon^2}{c_0c_2CSAH^2}\max\{\sigma^+,1/H\} \geq \frac{1}{4},\end{aligned} \tag{128}$$

as long as the sample size, $T = KH$, of the dataset is selected as

$$T \leq \frac{c_0c_2CSAH^3\min\{1/\sigma^+,H\}}{256(c_5)^2\varepsilon^2} \leq \frac{c_0c_2C_{\mathrm{r}}^\star SAH^3\min\{1/\sigma^+,H\}}{256(c_5)^2\varepsilon^2}. \tag{129}$$

E.2.3 Step 3: Putting all together

Next, we establish (108) by contradiction. Suppose that there exists an estimator $\widetilde{\mu}$ such that

$$\max_{(f,\phi\in\mathcal{F})\times\Phi} \mathbb{P}_{f,\phi}\left[\left\{V_1^{\star,\sigma^+,f,\phi}(\varrho) - V_1^{\widetilde{\mu},\sigma^+,f,\phi}(\varrho)\right\} \geq \varepsilon\right] < \frac{1}{4}. \tag{130}$$

According to (108), we would need

$$\forall w \in \mathcal{F}: \quad \max_{\phi\in\Phi} \mathbb{P}_{f,\phi}\left[\left\{V_1^{\star,\sigma^+,f,\phi}(m_f) - V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right\} \geq \varepsilon\right] < \frac{1}{4}. \tag{131}$$

To meet (131) for any $w \in \mathcal{F}$, we would require

$$\forall \phi \in \Phi : \mathbb{P}_{f,\phi} \left\{ V_1^{\star,\sigma^+,f,\phi}(m_f) - V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) < \varepsilon \right\} \geq \frac{3}{4}, \tag{132}$$

which, in view of (112), indicates that we would need

$$\forall \phi \in \Phi : \quad \mathbb{P}_{f,\phi} \left\{ \sum_{h=1}^{H} \left\| \widetilde{\mu}_h(\cdot \,|\, m_f) - \widetilde{\mu}_h^{\star,f,\phi}(\cdot \,|\, m_f) \right\|_1 < \frac{H}{8} \right\} \geq \frac{3}{4}. \tag{133}$$

On the other hand, (117) indicates

$$\forall \phi \in \Phi : \mathbb{P}_{f,\phi} \left[\widehat{\phi} = \phi \right] \geq \frac{3}{4}. \tag{134}$$

To achieve (130) or, in other words, (133), we apply (134) to all $w \in \mathcal{F}$, which would require

$$\forall (f,\phi) \in \mathcal{F} \times \Phi : \quad \mathbb{P}_{f,\phi} \left[(\widehat{f}, \widehat{\phi}) = (f,\phi) \right] \geq \frac{3}{4}. \tag{135}$$

This contract with (128) as long as the sample size condition in (129) is satisfied. Thus, if the sample size obeys (129), we cannot achieve an estimate $\widetilde{\mu}$ that satisfies (130), which completes the proof.

F Multiplayer General-sum Markov Games

In this section, we extend RTZ-VI-LCB to the setting of multi-player general-sum Markov games and present the corresponding theoretical guarantees.

F.1 Problem formulation

A robust general-sum Markov game is a tuple $\mathcal{M}(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, H, \{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{i=1}^m, \{r_i\}_{i=1}^m)$ with m players, where $\mathcal{S}$ denotes the state space and H is the horizon length. We have m different action spaces, where $\mathcal{A}_i$ is the action space for the i^{th} player and $|\mathcal{A}_i| = A_i$. We let $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ denote the joint action space, and let $\boldsymbol{a} := (a_1, \cdots, a_m) \in \mathcal{A}$ denote the joint actions by all m players. A notable deviation from standard MGs is that: for $1 \leq i \leq m$, instead of assuming a fixed transition kernel, each i^{th} player anticipates that the transition kernel is allowed to be chosen arbitrarily from a prescribed uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$. Here, the uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$ is constructed centered on $P^0(\cdot|s,\boldsymbol{a})$, with its size and shape defined by a certain distance metric ρ and a radius parameter $\sigma_i > 0$. $r_i = \{r_{h,i}\}_{h\in[H]}$ is a collection of reward functions for the i^{th} player, so that $r_{h,i}(s,\boldsymbol{a})$ gives the reward received by the i^{th} player if actions $\boldsymbol{a}$ are taken at state s at step h.

The policy of the i^{th} player is denoted as $\pi_i := \left\{\pi_{h,i} : \mathcal{S} \to \Delta_{\mathcal{A}_i}\right\}_{h\in[H]}$. We denote the product policy of all players as $\pi := \pi_1 \times \cdots \times \pi_M$, and denote the policy of all players except the i^{th} player as π_{-i}. We define $V_{h,i}^{\pi}(s)$ as the expected cumulative reward that will be received by the i^{th} player if starting at state s at step h and all players follow policy π. For any strategy π_{-i}, there also exists a *robust best response* of the i^{th} player, which is a policy $\mu^\star(\pi_{-i})$ satisfying $V_{h,i}^{\mu^\star(\pi_{-i}),\pi_{-i},\sigma_i}(s) = \sup_{\pi_i} V_{h,i}^{\pi_i,\pi_{-i},\sigma_i}(s)$ for any $(s,h) \in \mathcal{S}\times[H]$. For convenience, we denote $V_{h,i}^{\star,\pi_{-i},\sigma_i} := V_{h,i}^{\mu^\star(\pi_{-i}),\pi_{-i},\sigma_i}$. The Q-functions of the robust best response can be defined similarly.

Similar to the definition of behavior policy $(\mu^{\mathsf{n}}, \nu^{\mathsf{n}})$, we use the short-hand notation for the occupancy distribution w.r.t. the behavior policy $\pi^{\mathsf{n}} = (\pi_i^{\mathsf{n}}, \pi_{-i}^{\mathsf{n}})$ as: $\forall (h,s,\boldsymbol{a}) \in [H]\times\mathcal{S}\times\mathcal{A}$,

$$d_h^{\mathsf{n},P^0}(s) = d_h^{\pi^{\mathsf{n}},P^0}(s) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \varrho^{\mathsf{n}}, \pi^{\mathsf{n}}, P^0); \tag{136a}$$

$$d_h^{\mathsf{n},P^0}(s,\boldsymbol{a}) = d_h^{\pi^{\mathsf{n}},P^0}(s,\boldsymbol{a}) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \varrho^{\mathsf{n}}, \pi^{\mathsf{n}}, P^0)\, \pi^{\mathsf{n}}(\boldsymbol{a} \,|\, s). \tag{136b}$$

Similarly, for any product policy $\pi = (\pi_i, \pi_{-i})$, there is, $\forall (h,s,\boldsymbol{a}) \in [H]\times\mathcal{S}\times\mathcal{A}$

$$d_h^{\pi_i,\pi_{-i},P}(s) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \varrho, \pi, P); \tag{137a}$$

$$d_h^{\pi_i,\pi_{-i},P}(s,\boldsymbol{a}) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \varrho, \pi, P)\, \pi_{i,h}(a_i \,|\, s)\, \pi_{-i,h}(\boldsymbol{a}_{-i} \,|\, s). \tag{137b}$$

Therefore, the robust variant of standard solution concepts—robust NE for Robust multi-player general-sum MGs is introcuded as follows: A product policy π is considered a *robust NE* if

$$\forall(s) \in \mathcal{S}, \quad V_1^{\pi,\sigma_i}(s) = V_h^{\star,\pi_{-i},\sigma^+}(s). \tag{138}$$

A robust NE signifies that given the product policy (π) of the opponents, no player can enhance their outcome by deviating from their current policy unilaterally when each player accounts for the worst-case scenario within their uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$ for all $i = 1, 2, \cdots, m$.

Since finding exact robust equilibria can be complex and may not always be feasible, practitioners often seek approximate equilibria. In this context, a product policy $\pi \in \Delta(\mathcal{A})$ can be termed an *ε-robust NE* if

$$\mathrm{Gap}(\pi) \coloneqq \max\left\{\left\{V_{i,1}^{\star,\pi_{-i},\sigma_i}(\varrho) - V_{i,1}^{\pi,\sigma_i}(\varrho)\right\}_{i=1}^m\right\} \leq \varepsilon, \tag{139}$$

where

$$V_1^{\star,\pi_{-i},\sigma_i}(\varrho) = \mathbb{E}_{s\sim\varrho} V_1^{\star,\pi_{-i},\sigma_i}(s), \qquad \text{and} \qquad V_1^{\star,\sigma_i}(\varrho) = \mathbb{E}_{s\sim\varrho} V_1^{\star,\sigma_i}(s).$$

The existence of robust NE has been established for general divergence functions used in the uncertainty set in [5].

Goal With a dataset collected from the nominal environment, our objective is to find a solution among the ε-robust NEs for the robust multi-player general-sum MG $\mathcal{MG}_{\mathsf{r}}$ w.r.t. a specified uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$ around the nominal kernel, while minimizing the number of samples required under partial coverage of the state-action space.

F.2 Multi-RTZ-VI-LCB

Here we present the Multi-RTZ-VI-LCB algorithm in Algorithm 4, which is an extension of Algorithm 2 for multi-player general-sum Markov games.

According to the empirical frequencies of state transitions, we can naturally construct an empirical estimate $\widehat{P}^0 = \{\widehat{P}_h^0\}_{h=1}^H$ of P^0, where

$$\widehat{P}_h^0\left(s' \,|\, s, \boldsymbol{a}\right) = \begin{cases} \frac{1}{N_h(s,\boldsymbol{a})} \sum_{j=1}^N \mathbb{1}\left\{\left(s_j, \boldsymbol{a}_j, s_j'\right) = (s, \boldsymbol{a}, s')\right\}, & \text{if } N_h\left(s, \boldsymbol{a}\right) > 0; \\ \frac{1}{S}, & \text{if } N_h\left(s, \boldsymbol{a}\right) = 0, \end{cases} \tag{140}$$

$$\widehat{r}_{i,h}\left(s, \boldsymbol{a}\right) = \begin{cases} r_{i,h}\left(s, \boldsymbol{a}\right), & \text{if } N_h\left(s, \boldsymbol{a}\right) > 0; \\ 0, & \text{if } N_h\left(s, \boldsymbol{a}\right) = 0, \end{cases} \tag{141}$$

for any $(i, h, s, \boldsymbol{a}, s') \in [m] \times [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathcal{S}$. Besides, $N_h(s, \boldsymbol{a})$ represents the total number of sample transitions from $(s, \boldsymbol{a})$ at step h, and

$$N_h(s, \boldsymbol{a}) \coloneqq \sum_{j=1}^N \mathbb{1}\big\{(s_j, \boldsymbol{a}_j) = (s, \boldsymbol{a})\big\}. \tag{142}$$

Before the details of Multi-RTZ-VI-LCB, we extend Algorithm 1 as Algorithm 3, which reduces statistical dependencies and produces a distributionally equivalent dataset $\mathcal{D}_0$ with independent samples. Similar to Lemma 3.1, we present the following lemma concerning the dataset $\mathcal{D}_0$, whose proof is similar to the context in Appendix C.

Lemma F.1. *The dataset produced by the two-stage subsampling method is distributionally identical to $\mathcal{D}_0$ with probability at least $1 - 8\delta$, where $\{N_h(s, \boldsymbol{a})\}$ are independent of the sample transitions in $\mathcal{D}^0$ and obey: $\forall(h, s, \boldsymbol{a}) \in [H] \times \mathcal{S} \times \mathcal{A}$,*

$$N_h(s, \boldsymbol{a}) \geq \frac{K d_h^{\mathsf{n}}(s, \boldsymbol{a})}{8} - 5\sqrt{K d_h^{\mathsf{n}}(s, \boldsymbol{a}) \log\frac{KH}{\delta}}. \tag{143}$$

Based on Algorithm 4, we propose a model-based approach for solving robust multi-player general-sum MGs using an approximate $\widehat{P}^0$ for P^0, as summarized in Algorithm 4.

Algorithm 3 Two-stage subsampling for Multi-RTZ-VI-LCB.

input Dataset $\mathcal{D}$, probability δ.

1: **Step 1: Data Partitioning.** Split $\mathcal{D}$ into two equal-sized subsets, $\mathcal{D}^{\mathsf{m}}$ and $\mathcal{D}^{\mathsf{a}}$, each containing $K/2$ trajectories.

2: **Step 2: Defining Transition Bounds.** For step h and state s, denote the number of transitions from $\mathcal{D}^{\mathsf{m}}$ (resp. $\mathcal{D}^{\mathsf{a}}$) as $N_h^{\mathsf{m}}(s)$ (resp. $N_h^{\mathsf{a}}(s)$). Construct the trimmed count as:

$$N_h^{\mathsf{t}}(s) \coloneqq \max\left\{N_h^{\mathsf{a}}(s) - 10\sqrt{N_h^{\mathsf{a}}(s)\log\frac{HS}{\delta}},\, 0\right\}.$$

3: **Step 3: Generating Subsampled Dataset.** Randomly sample transitions (quadruples of the form $(s, \boldsymbol{a}, h, s')$) from $\mathcal{D}^{\mathsf{m}}$ uniformly. For each $(s,h) \in \mathcal{S}\times[H]$, include $\min\{N_h^{\mathsf{t}}(s), N_h^{\mathsf{m}}(s)\}$ transitions in the new dataset $\mathcal{D}^{\mathsf{t}}$.

output Set $\mathcal{D}_0 = \mathcal{D}^{\mathsf{t}}$.

Algorithm 4 Multi-RTZ-VI-LCB.

1: **Initialization**: Set uncertainty levels σ_i for $i = 1, 2, \cdots, m$; set $\widehat{V}_{i,h}^{\sigma_i}(s) = H$ and $\widehat{Q}_{i,h}^{\sigma_i}(s,\boldsymbol{a}) = H$ for all $(i, s, \boldsymbol{a}, h) \in [m]\times\mathcal{S}\times\mathcal{A}\times[H+1]$.

2: **Compute** the empirical reward function $\widehat{r}$ using (141) and the empirical transition kernel $\widehat{P}_0$ using (140).

3: **for** $h = H, H-1, \ldots, 1$ **do**

4: **Update** the robust Q-value estimate as

$$\widehat{Q}_{i,h}^{\sigma_i}(s,\boldsymbol{a}) = \min\left\{\widehat{r}_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}\left(\widehat{P}^0_{h,s,\boldsymbol{a}}\right)} P\widehat{V}_{i,h+1}^{\sigma_i} + \beta_{i,h}\left(s,\boldsymbol{a},\widehat{V}_{i,h+1}^{\sigma_i}\right),\, H\right\},$$

with $\beta_{i,h}(s,\boldsymbol{a},V) = \min\left\{\max\left\{\sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_h(s,\boldsymbol{a})}\mathsf{Var}_{\widehat{P}^0_{h,s,\boldsymbol{a}}}(V)}, \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_h(s,\boldsymbol{a})}\right\}, H\right\}$.

5: **Compute** Nash policy for each $s\in\mathcal{S}$ as

$$\pi_h(s) = (\pi_{i,h}(s), \pi_{-i,h}(s)) = \mathsf{ComputNash}\left(\widehat{Q}_{i,h}^{\sigma_i}(s,\cdot)\right),$$

6: **Update** the robust value estimate for each $s\in\mathcal{S}$ as

$$\widehat{V}_{i,h}^{\sigma_i}(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi_h(s)}\left[\widehat{Q}_{i,h}^{\sigma_i}(s,\boldsymbol{a})\right].$$

7: **end for**

output The product policy $\hat{\pi}(s) = \{\pi_h(s)\}_{h=1}^H$ with $\pi_h(s) = \prod_{i=1}^m \pi_{i,h}(s)$.

Similar to (16), we can tackle the multi-player general-sum MGs problem as:

$$\inf_{P\in\mathcal{U}^{\sigma_i}\left(\widehat{P}^0_{h,s,\boldsymbol{a}}\right)} P\widehat{V}_{i,h+1}^{\sigma_i} = \max_{\alpha\in[\min_s \widehat{V}_{i,h+1}^{\sigma_i}, \max_s \widehat{V}_{i,h+1}^{\sigma_i}]}\left\{\widehat{P}^0_{h,s,\boldsymbol{a}}\left[\widehat{V}_{i,h+1}^{\sigma_i}\right]_\alpha - \sigma_i\left(\alpha - \min_{s'}\left[\widehat{V}_{i,h+1}^{\sigma_i}\right]_\alpha(s')\right)\right\}. \tag{144}$$

where $\left[\widehat{V}_{i,h+1}^{\sigma_i}\right]_\alpha$ respectively denote the clipped versions of $\widehat{V}_{i,h+1}^{\sigma_i}\in\mathbb{R}^S$ based on some level $\alpha\geq 0$, as follows.

$$\left[\widehat{V}_{i,h+1}^{\sigma_i}\right]_\alpha(s) := \begin{cases}\widehat{V}_{i,h+1}^{\sigma_i}(s), & \text{if } \widehat{V}_{i,h+1}^{\sigma_i}(s) > \alpha;\\ \alpha. & \text{otherwise;}\end{cases} \tag{145}$$

F.3 Analysis of Multi-ME-Nash-QL

In this subsection, we prove Theorem 4.5, which can separated into three steps as the proof of Theorem 4.2.

First of all, similar to Assumption 4.1, we measure the distributional discrepancy between the historical data and the target data to assess the effectiveness of the historical dataset for achieving the desired goal. We propose a novel assumption for robust multi-agent general-sum MGs as:

Assumption F.2 (Robust multiple clipped concentrability)**.** The behavior policies of the historical dataset $\mathcal{D}$ satisfies

$$\max\left\{\left\{\sup_{(\pi_{-i},s,\boldsymbol{a},h,P)\in\Delta(\mathcal{A}_{-i})\times\mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{U}^{\sigma_i}(P^0)}\frac{\min\left\{d_h^{\pi_i^\star,\pi_{-i},P}(s,\boldsymbol{a}),\frac{1}{S\sum_{i=1}^m A_i}\right\}}{d_h^{\mathsf{n},P^0}(s,\boldsymbol{a})}\right\}_{i=1}^m\right\}\leq C_{\mathrm{mr}}^\star \tag{146}$$

F.3.1 Step 1: decoupling statistical dependency

Before bounding $\mathrm{Gap}(\widehat{\pi})$, we introduce an important lemma whose proof is similar to Lemma D.1 in Appendix G.1, quantifying the difference between $\widehat{P}$ and P when projected in the direction of the value function.

Lemma F.3. *Instate the assumptions in Theorem 4.5. Consider any vector $V\in\mathbb{R}^S$ with $\|V\|_\infty\leq H$ for all $(i,h,s,\boldsymbol{a})\in[m]\times[H]\times\mathcal{S}\times\mathcal{A}$ satisfying $N_h(s,\boldsymbol{a})>0$. With probability at least $1-\delta$, one has*

$$\left|\inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})}PV-\inf_{P\in\mathcal{U}^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})}PV\right|\leq C_4\sqrt{\frac{1}{N_h(s,\boldsymbol{a})}\mathsf{Var}_{\widehat{P}^0_{h,s,\boldsymbol{a}}}(V)\log\frac{KH}{\delta}}+C_4\frac{H\log\frac{KH}{\delta}}{N_h(s,\boldsymbol{a})} \tag{147}$$

for some sufficiently large constant $C_4>0$, and

$$\mathsf{Var}_{\widehat{P}^0_{h,s,\boldsymbol{a}}}(V)\leq 2\mathsf{Var}_{P^0_{h,s,\boldsymbol{a}}}(V)+O\left(\frac{H^2}{N_h(s,\boldsymbol{a})}\log\frac{KH}{\delta}\right). \tag{148}$$

With Lemma F.3, we can now have

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})}PV-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})}PV\right|\leq\beta_h(s,\boldsymbol{a},V) \tag{149}$$

for any $(i,h,s,\boldsymbol{a})\in[m]\times[H]\times\mathcal{S}\times\mathcal{A}$ satisfying $N_h(s,\boldsymbol{a})\geq 1$.

Therefore, we conclude that $\widehat{Q}_{i,h}^{\sigma_i}(s,\boldsymbol{a})$ is an optimistic estimation of $\widehat{Q}_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a})$ for any $i=1,2,\cdots,m$, which is summarized below, whose proof is similar to Lemma D.2 in Appendix G.2.

Lemma F.4. *With probability exceeding $1-\delta$, it holds that*

$$\widehat{Q}_{i,h}^{\sigma_i}(s,\boldsymbol{a})\geq Q_{i,h}^{\star,\widehat{\pi}_{-i},\sigma_i}(s,\boldsymbol{a})\qquad\text{and}\qquad\widehat{V}_{i,h}^{\sigma_i}(s)\geq V_{i,h}^{\star,\widehat{\pi}_{-i},\sigma_i}(s). \tag{150}$$

Besides, we introduce another key lemma highlighting the difference between robust multi-player general-sum MGs and standard multi-player general-sum MGs from the same idea of Lemma D.3, as shown below.

Lemma F.5. *Consider any multi-player general-sum MGs $\mathcal{MG}_{\mathsf{r}}=\left\{\mathcal{S},\{\mathcal{A}_i\}_{i=1}^m,H,\{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{i=1}^m,\{r_i\}_{i=1}^m\right\}$ and the uncertainty set $\{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{i=1}^m(\cdot)$ with TV distance. The optimistic robust value function estimate $\widehat{V}_{i,h}^{\sigma_i}$:*

$$\forall(i,h)\in[m]\times[H]:\quad\max_{s\in\mathcal{S}}\widehat{V}_{i,h}^{\sigma_i}-\min_{s\in\mathcal{S}}\widehat{V}_{i,h}^{\sigma_i}\leq\min\left\{\frac{(H+1)\left(1-(1-\sigma_i)^{H-h}\right)}{\sigma_i},H\right\}.$$

F.3.2 Step 2: decomposing the error $\mathrm{Gap}(\widehat{\pi})$

The goal of our algorithm is to output an ε-robust NE policy $(\widehat{\pi})$ satisfying $\mathrm{Gap}(\widehat{\pi})$ in (139), i.e.,

$$\mathrm{Gap}(\widehat{\pi}):=\max\left\{\left\{V_{i,1}^{\star,\widehat{\pi}_{-i},\sigma_i}(\varrho)-V_{i,1}^{\widehat{\pi},\sigma_i}(\varrho)\right\}_{i=1}^m\right\}\leq\varepsilon.$$

According to the relationship in Lemma F.4, under the definition of $\mathcal{A}_{-i} \coloneqq \mathcal{A}_1 \times \cdots \times \mathcal{A}_{i-1} \times \mathcal{A}_{i+1} \times \cdots \times \mathcal{A}_m$, we obtain

$$V_h^{\star,\widehat{\pi}_{-i,h},\sigma^+}(s) \le \widehat{V}_{i,h}^{\sigma_i}(s) = \min_{\pi_{-i} \in \Delta(\mathcal{A}_{-i})} \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}(s))} \left[Q_{i,h}^{\sigma_i}(s, \boldsymbol{a}) \right], \tag{151}$$

where the first equality comes from line 8 in Algorithm 4. Therefore, there exists a deterministic policy $\pi_{-i}^{\mathsf{d}} : \mathcal{S} \leftarrow \Delta(\mathcal{A}_{-i})$ satisfying that for any $s \in \mathcal{S}$

$$\pi_{-i}^{\mathsf{d}}(s) \ge \arg \min_{\pi_{-i} \in \Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}(s))} \left[Q_{i,h}^{\sigma_i}(s, \boldsymbol{a}) \right]. \tag{152}$$

Before starting, we introduce several useful notations:

- The state-action space covered by the behavior policy π^{n} in the nominal transition kernel P^0 is denoted as
$$\mathcal{C}^{\mathsf{n}} = \{(h, s, \boldsymbol{a}) : d_h^{\mathsf{n}}(s, \boldsymbol{a}) > 0\}. \tag{153}$$
- For any $(i, h) \in [m] \times [H]$, the set of potential state occupancy distributions w.r.t. the policy $(\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))$ and the uncertainty set $P \in \mathcal{U}^{\sigma_i}\left(P^0\right)$ f is denoted as
$$\mathcal{D}_{i,h}^{\mathsf{pi}} \coloneqq \left\{ \left[d_h^{\pi_i(s), \pi_{-i}^{\mathsf{b}}(s), P}(s) \right]_{s \in \mathcal{S}} : P \in \mathcal{U}^{\sigma_i}\left(P^0\right) \right\}; \tag{154}$$
$$\mathcal{D}_{i,h}^{\mathsf{pai}} \coloneqq \left\{ \left[d_h^{\pi_i(s), \pi_{-i}^{\mathsf{b}}(s), P}(s, \boldsymbol{a}) \right]_{(s,\boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}} : P \in \mathcal{U}^{\sigma_i}\left(P^0\right) \right\}. \tag{155}$$
- For convenience and without ambiguity, we introduce an additional notation for $(i, h) \in [m] \times [H]$ as
$$\beta_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}}(s) = \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))} \beta_{i,h}\left(s, \boldsymbol{a}, \widehat{V}_{i,h+1}^{\sigma_i}\right).$$
In particular, the vector $\beta_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}} \in \mathbb{R}^S$ is defined with its s-th item given by $\beta_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}}(s)$.
- Similarly, we can define the notation related to rewards for $(i, h) \in [m] \times [H]$ as
$$\widehat{r}_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}}(s) = \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))} \widehat{r}_{i,h}(s, \boldsymbol{a}).$$

To proceed with the analysis, we need to introduce a pessimistic V-estimation $\underline{V}_{i,h}^{\sigma_i}(s)$ that are defined similarly as $\widehat{V}_{i,h}^{\sigma_i}(s)$. Similar to Lemma F.4, we have $\underline{V}_{i,h}^{\sigma_i}(s) \le V_{i,h}^{\widehat{\pi}, \sigma_i}(s)$.

Therefore, according to the update rule in line 4 in Algorithm 4 and robust Bellman equality similar to (28), we derive

$$\begin{aligned}
&V_{i,h}^{\star,\widehat{\pi}_{-i},\sigma^+}(s) - V_{i,h}^{\widehat{\pi},\sigma^+}(s) \\
\le& \widehat{V}_{i,h}^{\sigma_i}(s) - \underline{V}_{i,h}^{\sigma_i}(s)(s) \\
\le& \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))} \inf_{P \in \mathcal{U}^{\sigma_i}\left(\widehat{P}_{h,s,\boldsymbol{a}}^0\right)} P \widehat{V}_{i,h+1}^{\sigma_i} + 2\beta_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}}(s) - \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))} \inf_{P \in \mathcal{U}^{\sigma_i}\left(P_{h,s,\boldsymbol{a}}^0\right)} P \underline{V}_{i,h+1}^{\sigma_i} \\
\le& \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))} \left[\inf_{P \in \mathcal{U}^{\sigma_i}\left(P_{h,s,\boldsymbol{a}}^0\right)} P \widehat{V}_{i,h+1}^{\sigma_i} - \inf_{P \in \mathcal{U}^{\sigma_i}\left(P_{h,s,\boldsymbol{a}}^0\right)} P \underline{V}_{i,h+1}^{\sigma_i} \right. \\
&\quad \left. + \left| \inf_{P \in \mathcal{U}^{\sigma_i}\left(P_{h,s,\boldsymbol{a}}^0\right)} P \widehat{V}_{i,h+1}^{\sigma_i} - \inf_{P \in \mathcal{U}^{\sigma_i}\left(\widehat{P}_{h,s,\boldsymbol{a}}^0\right)} P \widehat{V}_{i,h+1}^{\sigma_i} \right| \right] + 2\beta_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}}(s) \\
\overset{\text{(i)}}{\le}& \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))} \left[\inf_{P \in \mathcal{U}^{\sigma_i}\left(P_{h,s,\boldsymbol{a}}^0\right)} P \widehat{V}_{i,h+1}^{\sigma_i} - \inf_{P \in \mathcal{U}^{\sigma_i}\left(P_{h,s,\boldsymbol{a}}^0\right)} P \underline{V}_{i,h+1}^{\sigma_i} \right] + 3\beta_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}}(s) \\
\overset{\text{(ii)}}{\le}& \mathbb{E}_{\boldsymbol{a} \sim (\pi_i(s), \pi_{-i}^{\mathsf{b}}(s))} \left[P_{i,h,s,\boldsymbol{a}}^{\inf, V}\left(\widehat{V}_{i,h+1}^{\sigma_i} - \underline{V}_{i,h+1}^{\sigma_i} \right) \right] + 3\beta_{i,h}^{\pi_i, \pi_{-i}^{\mathsf{b}}}(s).
\end{aligned} \tag{156}$$

Here, (i) in (156) exists due to (149) in Lemma F.3 for $N_h(s,\boldsymbol{a})>0$ and

$$\left|\inf_{P\in\mathcal{U}^{\sigma_i}\left(P^0_{h,s,\boldsymbol{a}}\right)} P\widehat{V}^{\sigma_i}_{i,h+1} - \inf_{P\in\mathcal{U}^{\sigma_i}\left(\widehat{P}^0_{h,s,\boldsymbol{a}}\right)} P\widehat{V}^{\sigma_i}_{i,h+1}\right| \le H = \beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,h}(s) \text{ for } N_h(s,\boldsymbol{a})=0; \quad (157)$$

and (ii) is valid under the notation

$$P^{\inf,V}_{i,h,s,\boldsymbol{a}} \coloneqq \operatorname{argmin}_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,\boldsymbol{a}}\right)} P\underline{V}^{\sigma_i}_{i,h+1} \quad (158)$$

and consequently,

$$\inf_{P\in\mathcal{U}^{\sigma_i}\left(P^0_{h,s,\boldsymbol{a}}\right)} P\underline{V}^{\sigma_i}_{i,h+1} = P^{\inf,V}_{i,h,s,\boldsymbol{a}}\underline{V}^{\sigma_i}_{i,h+1}, \text{ and } \inf_{P\in\mathcal{U}^{\sigma_i}\left(P^0_{h,s,\boldsymbol{a}}\right)} P\widehat{V}^{\sigma_i}_{i,h+1} \le P^{\inf,V}_{i,h,s,\boldsymbol{a}}\widehat{V}^{\sigma_i}_{i,h+1}.$$

For ease of proof, we introduce a notation as $\check{P}^{\inf,V}_{i,h,s} \coloneqq \mathbb{E}_{\boldsymbol{a}\sim(\pi_i(s),\pi^{\mathsf{b}}_{-i}(s))}P^{\inf,V}_{i,h,s,\boldsymbol{a}}$. Furthermore, we define a sequence of matrices $\check{P}^{\inf,V}_{i,h}\in\mathbb{R}^{S\times S}$. We can utilizing (156) recursively over the time steps $h,h+1,\cdots,H$ and derive

$$\begin{aligned}
V^{\star,\widehat{\pi}_{-i},\sigma^+}_{i,h}(s) - V^{\widehat{\pi},\sigma^+}_{i,h}(s) &\le \widehat{V}^{\sigma_i}_{i,h}(s) - \underline{V}^{\sigma_i}_{i,h}(s) \\
&\le \check{P}^{\inf,V}_{i,h}\left(\widehat{V}^{\sigma_i}_{i,h+1} - \underline{V}^{\sigma_i}_{i,h+1}\right) + 3\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,h}(s) \\
&\le \check{P}^{\inf,V}_{i,h}\check{P}^{\inf,V}_{i,h+1}\left(\widehat{V}^{\sigma_i}_{i,h+2} - \underline{V}^{\sigma_i}_{i,h+2}\right) + 3\check{P}^{\inf,V}_{i,h}\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,h+1} + 3\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,h}(s) \\
&\le \cdots \le 3\sum_{i'=h}^{H}\left(\prod_{j=h+1}^{i'-1}\check{P}^{\inf,V}_{i,j}\right)\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,i'} + 3\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,h}(s) \\
&= 3\sum_{i'=h}^{H}\left(\prod_{j=h}^{i'-1}\check{P}^{\inf,V}_{i,j}\right)\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,i'}, \qquad (159)
\end{aligned}$$

where we define $\left(\prod_{j=h}^{i'-1}\check{P}^{\inf,V}_{i,j}\right) = I$ for conciseness.

For any $d^{\pi_i,\pi^{\mathsf{b}}_{-i}}_h \in \mathcal{D}^{\mathsf{pi}}_h$ (cf. (53)), taking inner product with (59) yields

$$\begin{aligned}
\left\langle d^{\pi_i,\pi^{\mathsf{b}}_{-i}}_h, V^{\star,\widehat{\pi}_{-i},\sigma^+}_{i,h} - V^{\widehat{\pi},\sigma^+}_{i,h}\right\rangle &\le \left\langle d^{\pi_i,\pi^{\mathsf{b}}_{-i}}_h, 3\sum_{i'=h}^{H}\left(\prod_{j=h}^{i'-1}\check{P}^{\inf,V}_{i,j}\right)\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,i'}\right\rangle \\
&= 3\sum_{i'=h}^{H}\left\langle d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}, \beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,i'}\right\rangle, \qquad (160)
\end{aligned}$$

where

$$d^{\mathsf{p},\pi^{\mathsf{d}}_i,\pi^{\star}_{-i}}_{i'} \coloneqq \left[\left(d^{\pi_i,\pi^{\mathsf{b}}_{-i}}_h\right)^\top\left(\prod_{j=h}^{i'-1}\check{P}^{\inf,V}_{i,j}\right)\right]^\top \in \mathcal{D}^{\mathsf{pi}}_{i'} \quad (161)$$

by the definition of $\mathcal{D}^{\mathsf{pi}}_{i'}$ (cf. (154)) for all $i'=h+1,\cdots,H$.

Next, we control $\langle d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}, \beta_{i,i'}^{\pi_i,\pi_{-i}^{\mathsf{b}}} \rangle$ utilizing concentrability. First of all, according to the definition of penalty, we demonstrate that the pessimistic penalty satisfies

$$\begin{aligned}
\beta_{i,i'}(s,\boldsymbol{a},\hat{V}) &\leq \max\left\{\sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})}\mathsf{Var}_{\widehat{P}^0_{i,s,\boldsymbol{a}}}(\widehat{V})}, \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})}\right\} \\
&\leq \sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})}\mathsf{Var}_{\widehat{P}^0_{i,s,\boldsymbol{a}}}(\widehat{V})} + \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})} \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})}\left(2\mathsf{Var}_{P^0_{i,s,\boldsymbol{a}}}(\widehat{V}) + \frac{C_0H^2}{N_i(s,\boldsymbol{a})}\log\frac{KH}{\delta}\right)} + \frac{2C_{\mathsf{n}}H\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{2C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})}\mathsf{Var}_{P^0_{i,s,\boldsymbol{a}}}(\widehat{V})} + \frac{\left(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_0}\right)H\log\frac{KH}{\delta}}{N_i(s,\boldsymbol{a})}
\end{aligned} \tag{162}$$

where (i) holds by applying (148) for some sufficiently large C_0 and (ii) exists follows from the Cauchy-Schwarz inequality. Therefore, combining the definition of $\beta_{i,i'}^{\pi_i,\pi_{-i}^{\mathsf{b}}}(s)$, we obtain

$$\begin{aligned}
\langle d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}, \beta_{i,i'}^{\pi_i,\pi_{-i}^{\mathsf{b}}} \rangle &= \sum_{s\in\mathcal{S}} d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}(s)\beta_{i,i'}^{\pi_i,\pi_{-i}^{\mathsf{b}}}(s) \\
&= \sum_{s\in\mathcal{S}} d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}(s)\mathbb{E}_{\boldsymbol{a}\sim(\pi_i(s),\pi_{-i}^{\mathsf{b}}(s))}\beta_{i,i'}(s,\boldsymbol{a},\hat{V}) \\
&= \sum_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}(s)\,\mathbb{1}\{a_i=\pi_i^\star(s)\}\pi_{-i}^{\mathsf{d}}(\mathbf{a}_{-i}|s)\beta_{i,i'}(s,\boldsymbol{a},\hat{V}) \\
&= \sum_{(s,a_i)\in\mathcal{S}\times\mathcal{A}} d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}(s,a_i,\pi_{-i}^{\mathsf{b}}(s))\beta_{i,i'}(s,\pi_i^{\mathsf{d}}(s),\mathbf{a}_{-i},\hat{V}),
\end{aligned} \tag{163}$$

where the last equation holds due to the definition in (137b). Then, we observe $d_h^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}(s,\boldsymbol{a}) \in \mathcal{D}_h^{\mathsf{pai}}$ (cf. (155)). Thereafter, we divide the bound (163) into two cases.

For the first case, i.e., $s \in S$ where $\max_{P\in\mathcal{U}^{\sigma_i}(P^0)} d_{i'}^{\pi_i,\pi_{-i}^{\mathsf{b}},P}\big(s,a_i,\pi_{-i}^{\mathsf{b}}(s)\big) = 0$, it follows from the definition (cf. (154)) that for any $d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}(s,a_i,\pi_{-i}^{\mathsf{b}}(s)) \in \mathcal{D}_i^{\mathsf{pai}}$, it satisfies that

$$d_{i'}^{\mathsf{p},\pi_i,\pi_{-i}^{\mathsf{b}}}(s,a_i,\pi_{-i}^{\mathsf{b}}(s)) = 0. \tag{164}$$

For the second case, i.e., $s \in S$ where $\max_{P\in\mathcal{U}^{\sigma^+}(P^0)} d_{i'}^{\pi_i,\pi_{-i}^{\mathsf{b}},P}\big(s,a_i,\pi_{-i}^{\mathsf{b}}(s)\big) > 0$, by the assumption in (146)

$$\max_{P\in\mathcal{U}^{\sigma_i}(P^0)} \frac{\min\left\{d_{i'}^{\pi_i,\pi_{-i}^{\mathsf{b}},P}\big(s,a_i,\pi_{-i}^{\mathsf{b}}(s)\big), \frac{1}{S\sum_{i=1}A_i}\right\}}{d_{i'}^{\mathsf{n}}\big(s,a_i,\pi_{-i}^{\mathsf{b}}(s)\big)} \leq C_{\mathsf{r}}^\star < \infty.$$

It implies that

$$d_{i'}^{\mathsf{n}}\big(s,a_i,\pi_{-i}^{\mathsf{b}}(s)\big) > 0 \quad \text{and} \quad \big(i',s,a_i,\pi_{-i}^{\mathsf{b}}(s)\big) \in \mathcal{C}^{\mathsf{n}}. \tag{165}$$

Lemma F.1 tells that with probability at least $1-8\delta$,

$$\begin{aligned}
N_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big) &\geq \frac{Kd^{\mathsf{n}}_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big)}{8}-5\sqrt{Kd^{\mathsf{n}}_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big)\log\frac{KH}{\delta}}\\
&\overset{\text{(i)}}{\geq} \frac{Kd^{\mathsf{n}}_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big)}{16}\\
&\overset{\text{(ii)}}{\geq} \frac{K\max_{P\in\mathcal{U}^{\sigma_i}(P^0)}\min\left\{d^{\pi_i,\pi^{\mathsf{b}}_{-i},P}_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big),\frac{1}{S\sum_{i=1}A_i}\right\}}{16C^\star_{\mathsf{r}}}\\
&\geq \frac{K\min\left\{d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s)),\frac{1}{S\sum_{i=1}A_i}\right\}}{16C^\star_{\mathsf{r}}},
\end{aligned}\tag{166}$$

where (ii) comes from Assumption F.2 and (i) holds due to

$$\begin{aligned}
Kd^{\mathsf{n}}_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big) &\geq c_0\frac{HS\sum_{i=1}A_i}{d^{\mathsf{n}}_{\mathsf{m}}}\log\frac{KH}{\delta}f(\{\sigma_i\}_{i=1}^m,H)d^{\mathsf{n}}_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big)\\
&\geq c_0HS\sum_{i=1}A_i\log\frac{KH}{\delta}f(\{\sigma_i\}_{i=1}^m,H)\geq 1600\log\frac{KH}{\delta},
\end{aligned}\tag{167}$$

where $f(\{\sigma_i\}_{i=1}^m,H)=\min\left\{\left\{\frac{(H\sigma_i-1+(1-\sigma_i)^H)}{(\sigma_i)^2}\right\}_{i=1}^m,H\right\}$, the first inequality follows from condition (26), and the second inequality follows from

$$d^{\mathsf{n}}_{\mathsf{m}}=\min_{h,s,a_i,\pi^{\mathsf{b}}_{-i}(s)}\big\{d^{\mathsf{n}}_h(s,\pi^{\mathsf{d}}_i(s),\mathbf{a}_{-i}):d^{\mathsf{n}}_h(s,\pi^{\mathsf{d}}_i(s),\mathbf{a}_{-i})>0\big\}\leq d^{\mathsf{n}}_{i'}\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big).\tag{168}$$

Combining the results in (63) and (64), we arrive at

$$\begin{aligned}
\langle d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'},\beta^{\pi_i,\pi^{\mathsf{b}}_{-i}}_{i,i'}\rangle &= \sum_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s))\beta_{i,i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s),\hat{V})\\
&\leq \sum_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s))\sqrt{\frac{2C_{\mathsf{n}}\log\frac{KH}{\delta}}{N_i\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big)}\mathsf{Var}_{P^0_{i,s,a_i,\pi^{\mathsf{b}}_{-i}(s)}}\big(\widehat{V}\big)}\\
&\quad+\sum_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s))\frac{\big(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_0}\big)H\log\frac{KH}{\delta}}{N_i\big(s,a_i,\pi^{\mathsf{b}}_{-i}(s)\big)}\\
&\leq \sum_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s))\Bigg(\frac{16C^\star_{\mathsf{r}}\big(2C_{\mathsf{n}}+\sqrt{C_{\mathsf{n}}C_0}\big)H\log\frac{KH}{\delta}}{K\min\left\{d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s)),\frac{1}{S\sum_{i=1}A_i}\right\}}\\
&\quad+\sqrt{\frac{32C^\star_{\mathsf{r}}C_{\mathsf{n}}\log\frac{KH}{\delta}}{K\min\left\{d^{\mathsf{p},\pi_i,\pi^{\mathsf{b}}_{-i}}_{i'}(s,a_i,\pi^{\mathsf{b}}_{-i}(s)),\frac{1}{S\sum_{i=1}A_i}\right\}}\mathsf{Var}_{P^0_{i,s,a_i,\pi^{\mathsf{b}}_{-i}(s)}}\big(\widehat{V}\big)}\Bigg).
\end{aligned}\tag{169}$$

Similar to the proof in Appendix D.2, we are ready to bound $V^{\star,\widehat{\pi}_{-i},\sigma^+}_{i,h}(\varrho)-V^{\widehat{\pi},\sigma^+}_{i,h}(\varrho)$. There exists some sufficiently large constants $C_1,C_2,C_3>0$, and

$$\begin{aligned}
V^{\star,\widehat{\pi}_{-i},\sigma^+}_{i,h}(\varrho)-V^{\widehat{\pi},\sigma^+}_{i,h}(\varrho) \leq&\sqrt{\frac{C^\star_{\mathsf{r}}C_1H^3S\sum_{i=1}A_i\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma_i-1+(1-\sigma_i)^H)}{(\sigma_i)^2},H\right\}}\\
&+\frac{C^\star_{\mathsf{r}}C_2H^2S\sum_{i=1}A_i\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma_i-1+(1-\sigma_i)^H)}{(\sigma_i)^2},H\right\}\\
\leq&\sqrt{\frac{C^\star_{\mathsf{r}}C_3H^3S\sum_{i=1}A_i\log\frac{KH}{\delta}}{K}\min\left\{\frac{2(H\sigma_i-1+(1-\sigma_i)^H)}{(\sigma_i)^2},H\right\}},
\end{aligned}\tag{170}$$

where the last inequality follows from condition (26).

F.3.3 Step 3: summing up the results

Consequently, we obtain the upper bound of $V_{i,1}^{\star,\widehat{\pi}_{-i},\sigma_i}(\varrho) - V_{i,1}^{\widehat{\pi},\sigma_i}(\varrho)$ in (170). which directly leads to

$$\mathrm{Gap}(\widehat{\pi}) \leq c_1 \sqrt{\frac{C_{\mathsf{r}}^{\star} H^2 S \sum_{i=1}^m A_i \log \frac{KH}{\delta}}{K} \min\left\{ \left\{ \frac{2(H\sigma_i - 1 + (1-\sigma_i)^H)}{(\sigma_i)^2} \right\}_{i=1}^m, H \right\}}, \quad (171)$$

for some sufficiently large c_1 and

$$K \geq HS \sum_{i=1} A_i \log \frac{KH}{\delta} \min\left\{ \left\{ \frac{2(H\sigma_i - 1 + (1-\sigma_i)^H)}{(\sigma_i)^2} \right\}_{i=1}^m, H \right\}.$$

G Proofs of lemmas for Theorem 4.2

G.1 Proof of Lemma D.1

We prove Lemma D.1 similar to the proof of Claim 1 introduced by [43].

G.1.1 Proof of (46).

According to the definition in (16), for any fixed value vector V independent of $\widehat{P}^0_{h,s,a,b}$, we have

$$\begin{aligned}
&\left| \inf_{P \in \mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} PV - \inf_{P \in \mathcal{U}^{\sigma^+}(P^0_{h,s,a,b})} PV \right| \\
=& \Big| \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ \widehat{P}^0_{h,s,a,b} [V]_\alpha - \sigma^+ \left(\alpha - \min_{s'} [V]_\alpha (s') \right) \right\} \\
&- \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P^0_{h,s,a,b} [V]_\alpha - \sigma^+ \left(\alpha - \min_{s'} [V]_\alpha (s') \right) \right\} \Big| \\
\leq& \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left| \widehat{P}^0_{h,s,a,b} [V]_\alpha - P^0_{h,s,a,b} [V]_\alpha \right| \\
\leq& \max_{\alpha \in [0,H]} \left| \widehat{P}^0_{h,s,a,b} [V]_\alpha - P^0_{h,s,a,b} [V]_\alpha \right|, \qquad (172)
\end{aligned}$$

where the last inequality exists due to the fact that the maximum operator is 1-Lipschitz. According to the definition of empirical transition kernel $\widehat{P}^0_{h,s,a,b}$, we have

$$\begin{aligned}
&\left(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b} \right) [V]_\alpha \\
=& \sum_{s' \in \mathcal{S}} \underbrace{[V(s')]_\alpha \left[\frac{\sum_{i=1}^N \mathbb{1}\{h_i = h, s_i = s, a_i = a, b_i = b, s_i' = s'\}}{N_h(s,a,b)} - P_h^0(s' \mid s,a,b) \right]}_{=: X_{s'}}
\end{aligned}$$

as a sum of independent random variables. Based on the relationship between $P^0_{h,s,a,b}$ and $\widehat{P}^0_{h,s,a,b}$, we verify $\mathbb{E}[X_{s'}] = 0$ and $|X_{s'}| \leq H$ for all $s' \in \mathcal{S}$. With probability exceeding $1 - \delta$ and for some universal constant $C_4 > 0$, under the Bernstein inequality [37, Theorem 2.8.4], we have

$$\begin{aligned}
\left(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b} \right) [V]_\alpha &\leq C_4 \sqrt{\frac{1}{N_h(s,a,b)} \mathsf{Var}_{P^0_{h,s,a,b}} ([V]_\alpha) \log \frac{KH}{\delta}} + \frac{C_4 H \log \frac{KH}{\delta}}{N_h(s,a,b)} \\
&\leq C_4 \sqrt{\frac{1}{N_h(s,a,b)} \mathsf{Var}_{P^0_{h,s,a,b}} (V) \log \frac{KH}{\delta}} + \frac{C_4 H \log \frac{KH}{\delta}}{N_h(s,a,b)}, \qquad (173)
\end{aligned}$$

where the last inequality comes from the definition of $[V]_\alpha$ in (17).

Let $\overline{V} \coloneqq V - \left(P^0_{h,s,a,b}V\right)1$, we have

$$\begin{aligned}
\mathsf{Var}_{P^0_{h,s,a,b}}(V) &= P^0_{h,s,a,b}\big(\overline{V}\circ\overline{V}\big) \\
&= \widehat{P}^0_{h,s,a,b}\big(\overline{V}\circ\overline{V}\big) + \big(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\big)\big(\overline{V}\circ\overline{V}\big) \\
&= \mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) + \big[\big(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\big)V\big]^2 + \big(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\big)\big(\overline{V}\circ\overline{V}\big),
\end{aligned} \tag{174}$$

where the last equation holds since

$$\begin{aligned}
\widehat{P}^0_{h,s,a,b}\big(\overline{V}\circ\overline{V}\big) &= \widehat{P}^0_{h,s,a,b}\left(\left[V - \left(P^0_{h,s,a,b}V\right)1\right]\circ\left[V - \left(P^0_{h,s,a,b}V\right)1\right]\right) \\
&= \widehat{P}^0_{h,s,a,b}\left(V\circ V\right) - 2\left(P^0_{h,s,a,b}V\right)\left(\widehat{P}^0_{h,s,a,b}V\right) + \left(P^0_{h,s,a,b}V\right)^2 \\
&= \widehat{P}^0_{h,s,a,b}\left(\left[V - \left(\widehat{P}^0_{h,s,a,b}V\right)1\right]\circ\left[V - \left(\widehat{P}^0_{h,s,a,b}V\right)1\right]\right) + \left(\widehat{P}^0_{h,s,a,b}V\right)^2 \\
&\quad - 2\left(P^0_{h,s,a,b}V\right)\left(\widehat{P}^0_{h,s,a,b}V\right) + \left(P^0_{h,s,a,b}V\right)^2 \\
&= \mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) + \big[\big(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\big)V\big]^2.
\end{aligned}$$

Analogous to (173), with probability exceeding $1-\delta$,

$$\begin{aligned}
\left|\big(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b}\big)\big(\overline{V}\circ\overline{V}\big)\right| &\le C_4\sqrt{\frac{1}{N_h(s,a,b)}\mathsf{Var}_{P^0_{h,s,a,b}}\big(\overline{V}\circ\overline{V}\big)\log\frac{KH}{\delta}} + \frac{C_4H^2\log\frac{KH}{\delta}}{N_h(s,a,b)} \\
&\le C_4\sqrt{\frac{H^2}{N_h(s,a,b)}\mathsf{Var}_{P^0_{h,s,a,b}}(V)\log\frac{KH}{\delta}} + \frac{C_4H^2\log\frac{KH}{\delta}}{N_h(s,a,b)},
\end{aligned} \tag{175}$$

where the last inequation comes from the fact that

$$\mathsf{Var}_{P^0_{h,s,a,b}}\big(\overline{V}\circ\overline{V}\big) \le P^0_{h,s,a,b}\big(\overline{V}\circ\overline{V}\circ\overline{V}\circ\overline{V}\big) \le H^2P^0_{h,s,a,b}\big(\overline{V}\circ\overline{V}\big) = H^2\mathsf{Var}_{P^0_{h,s,a,b}}(V).$$

Employing (175), we further bound (174) as

$$\begin{aligned}
\mathsf{Var}_{P^0_{h,s,a,b}}(V) \le& \mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) + \big[\big(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\big)V\big]^2 \\
&+ C_4\sqrt{\frac{H^2\log\frac{KH}{\delta}}{N_h(s,a,b)}\mathsf{Var}_{P^0_{h,s,a,b}}(V)} + \frac{C_4H^2\log\frac{KH}{\delta}}{N_h(s,a,b)} \\
\le& \mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) + \big[\big(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\big)V\big]^2 + \frac{C_4H^2\log\frac{KH}{\delta}}{N_h(s,a,b)} \\
&+ \frac{1}{2}\mathsf{Var}_{P^0_{h,s,a,b}}(V) + \frac{C_4^2H^2\log\frac{KH}{\delta}}{2N_h(s,a,b)},
\end{aligned}$$

where the last relation holds due to the AM-GM inequality. Therefore, we obtain

$$\mathsf{Var}_{P^0_{h,s,a,b}}(V) \le 2\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) + 2\big[\big(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\big)V\big]^2 + \frac{\left(C_4^2+2C_4\right)H^2\log\frac{KH}{\delta}}{N_h(s,a,b)}. \tag{176}$$

Combining (176) and (173), we derive

$$\begin{aligned}
\left|\big(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b}\big)V\right| \le& \sqrt{\frac{2C_4^2}{N_h(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V)\log\frac{KH}{\delta}} + \frac{\sqrt{C_4^2\left(C_4^2+2C_4\right)H\log\frac{KH}{\delta}}}{N_h(s,a,b)} \\
&+ \sqrt{\frac{2C_4^2}{N_h(s,a,b)}\log\frac{KH}{\delta}}\left|\big(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b}\big)V\right| + \frac{C_4H\log\frac{KH}{\delta}}{N_h(s,a,b)}.
\end{aligned} \tag{177}$$

Next, we consider two cases, i.e., $N_h(s,a,b) \leq \frac{1}{8C_4^2}\log\frac{KH}{\delta}$ and $N_h(s,a,b) > \frac{1}{8C_4^2}\log\frac{KH}{\delta}$. In the first case of $N_h(s,a,b) \leq \frac{1}{8C_4^2}\log\frac{KH}{\delta}$, (46) is valid since

$$\left|\inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} PV - \inf_{P\in\mathcal{U}^{\sigma^+}(P^0_{h,s,a,b})} PV\right| \leq \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left|\left(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b}\right)V\right| \leq 2H = O\left(\frac{H\log\frac{KH}{\delta}}{N_h(s,a,b)}\right). \tag{178}$$

In the second case of $N_h(s,a,b) > \frac{1}{8C_4^2}\log\frac{KH}{\delta}$, it follows from (177) that

$$\left|\left(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b}\right)V\right| \leq \frac{1}{2}\left|\left(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b}\right)V\right| + \sqrt{\frac{2C_4^2}{N_h(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V)\log\frac{KH}{\delta}} + \frac{C_4+\sqrt{C_4^2(C_4^2+2C_4)}}{N_h(s,a,b)}H\log\frac{KH}{\delta},$$

which, by arranging the terms, leads to

$$\left|\left(\widehat{P}^0_{h,s,a,b} - P^0_{h,s,a,b}\right)V\right| \leq \sqrt{\frac{8C_4^2}{N_h(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V)\log\frac{KH}{\delta}} + 2H\frac{C_4+\sqrt{C_4^2(C_4^2+2C_4)}}{N_h(s,a,b)}\log\frac{KH}{\delta}. \tag{179}$$

Combining (172) and (179), we obtain

$$\left|\inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} PV - \inf_{P\in\mathcal{U}^{\sigma^+}(P^0_{h,s,a,b})} PV\right| \leq \sqrt{\frac{8C_4^2}{N_h(s,a,b)}\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V)\log\frac{KH}{\delta}} + 2H\frac{C_4+\sqrt{C_4^2(C_4^2+2C_4)}}{N_h(s,a,b)}\log\frac{KH}{\delta}. \tag{180}$$

Joining these two cases, we conclude the proof of (46).

G.1.2 Proof of (47).

We consider two cases, i.e., $N_h(s,a,b) < 16C_4^2\log\frac{KH}{\delta}$ and $N_h(s,a,b) \geq 16C_4^2\log\frac{KH}{\delta}$. In the first case of $N_h(s,a,b) < 16C_4^2\log\frac{KH}{\delta}$, (47) is valid since

$$\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) \leq H^2 = O\left(\frac{H^2\log\frac{KH}{\delta}}{N_h(s,a,b)}\right).$$

In the second case of $N_h(s,a,b) \geq 16C_4^2\log\frac{KH}{\delta}$, we have

$$\begin{aligned}
\mathsf{Var}_{\widehat{P}^0_{h,s,a,b}}(V) &\overset{\text{(i)}}{=} \mathsf{Var}_{P^0_{h,s,a,b}}(V) - \left[\left(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\right)V\right]^2 - \left(P^0_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\right)\left(\overline{V}\circ\overline{V}\right) \\
&\overset{\text{(ii)}}{\leq} \mathsf{Var}_{P^0_{h,s,a,b}}(V) + C_4\sqrt{\frac{H^2}{N_h(s,a,b)}\mathsf{Var}_{P^0_{h,s,a,b}}\left(V\right)\log\frac{KH}{\delta}} + \frac{C_4H^2\log\frac{KH}{\delta}}{N_h(s,a,b)} \\
&\overset{\text{(iii)}}{\leq} 2\mathsf{Var}_{P^0_{h,s,a,b}}(V) + \frac{\left(C_4^2/4+C_4\right)H^2\log\frac{KH}{\delta}}{N_h(s,a,b)} \\
&= 2\mathsf{Var}_{P^0_{h,s,a,b}}(V) + O\left(\frac{H^2\log\frac{KH}{\delta}}{N_h(s,a,b)}\right),
\end{aligned}$$

where (i) comes from (174), (ii) holds due to (175), and (iii) is based on the AM-GM inequality. Combining the two cases, (47) holds and Lemma D.1 is proved.

G.2 Proof of Lemma D.2

Assuming that $\widehat{Q}_h^+(s,a,b) \geq Q_h^{\star,\widehat{\nu},\sigma^+}(s,a,b)$, then we can obtain $\widehat{V}_h^+(s) \geq V_h^{\star,\nu,\sigma^+}(s)$, since

$$\begin{aligned}\widehat{V}_h^+(s) &= \mathbb{E}_{a\sim\mu_h^+(s),b\sim\nu_h^+(s)}\left[\widehat{Q}_h^+(s,a,b)\right] \\ &\overset{\text{(i)}}{\geq} \mathbb{E}_{a\sim\mu^\star(s),b\sim\widehat{\nu}(s)}\left[\widehat{Q}_h^+(s,a,b)\right] \geq \mathbb{E}_{a\sim\mu^\star(s),b\sim\widehat{\nu}(s)}\left[Q_h^{\star,\widehat{\nu},\sigma^+}(s,a,b)\right] = V_h^{\star,\widehat{\nu},\sigma^+}(s),\end{aligned}$$

where (i) holds due to the fact that $\widehat{\nu} = \nu_h^+$ and (μ_h^+, ν_h^+) is the Nash equilibrium of $\widehat{Q}_h^+(s,a,b)$. Hence, we need to verify

$$\widehat{Q}_h^+(s,a,b) \geq Q_h^{\star,\widehat{\nu},\sigma^+}(s,a,b), \tag{181}$$

which can be proved by mathematical induction. Specifically, (181) holds when $h = H+1$ under the trivial fact $\widehat{Q}_{H+1}^+(s,a,b) = Q_{H+1}^{\star,\sigma^+}(s,a,b) = 0$. Suppose that (181) holds for all $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ at some time step $h \in [H]$. According to the update rule in line 4 in Algorithm 2, (181) exists if $\widehat{Q}_h^+(s,a,b) = H$ because $\widehat{Q}_h^+(s,a,b) = H \geq Q_h^{\star,\widehat{\nu},\sigma^+}(s,a,b)$. Besides, in the case of $N_h(s,a,b) = 0$, we have $\beta_h\left(s,a,b,\widehat{V}_{h+1}^+\right) = H$, leading to $\widehat{Q}_h^+(s,a,b) = H \geq Q_h^{\star,\widehat{\nu},\sigma^+}(s,a,b)$. Otherwise, for $N_h(s,a,b) > 0$, $\widehat{Q}_h^+(s,a,b)$ is updated as

$$\begin{aligned}\widehat{Q}_h^+(s,a,b) =& \widehat{r}(s,a,b) + \inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ + \beta_h\left(s,a,b,\widehat{V}_{h+1}^+\right) \\ \geq& \widehat{r}(s,a,b) + \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ + \beta_h\left(s,a,b,\widehat{V}_{h+1}^+\right) \\ &- \left|\inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ - \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+\right| \\ \geq& \widehat{r}(s,a,b) + \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^+ + 0 \\ \geq& \widehat{r}(s,a,b) + \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{h,s,a,b}\right)} P\widehat{V}_{h+1}^{\star,\widehat{\nu},\sigma^+} + 0 = Q_h^{\star,\widehat{\nu},\sigma^+}(s,a,b),\end{aligned} \tag{182}$$

where the second inequality holds due to (48) in Lemma D.1 and the last equality comes from the empirical robust Bellman equation (30). Together with the case of $h = H+1$, we complete prove Lemma D.2.

G.3 Proof of Lemma D.3

Following the proof by Lemma 3 in [34], we bound $\min_{s\in\mathcal{S}}\widehat{V}_h^+(s)$ and $\max_{s\in\mathcal{S}}\widehat{V}_h^+(s)$. Specifically,

$$\begin{aligned}\min_{s\in\mathcal{S}}\widehat{V}_h^+(s) &= \min_{s\in\mathcal{S}}\mathbb{E}_{(a,b)\sim\mu_h^+\times\nu_h^+}\left[\widehat{Q}_h^+(s,a,b)\right] \\ &= \min_{s\in\mathcal{S}}\mathbb{E}_{(a,b)\sim\mu_h^+\times\nu_h^+}\left[\widehat{r}_h(s,a,b) + \inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} P\widehat{V}_{h+1}^+ + \beta_h\left(s,a,b,\widehat{V}_{h+1}\right)\right] \\ &\geq 0 + \min_{s\in\mathcal{S}}\widehat{V}_{h+1}^+(s) + 0,\end{aligned} \tag{183}$$

where the second equality is valid due to the update rule in line 4 in Algorithm 2. Similarly,

$$\begin{aligned}\max_{s\in\mathcal{S}}\widehat{V}_h^+ &= \max_{s\in\mathcal{S}}\mathbb{E}_{(a,b)\sim\mu_h^+\times\nu_h^+}\left[\widehat{Q}_h^+(s,a,b)\right] \\ &= \max_{s\in\mathcal{S}}\mathbb{E}_{(a,b)\sim\mu_h^+\times\nu_h^+}\left[\widehat{r}_h(s,a,b) + \inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} P\widehat{V}_{h+1}^+ + \beta_h\left(s,a,b,\widehat{V}_{h+1}\right)\right] \\ &\leq 1 + \max_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}}\inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}_{h,s,a,b})} P\widehat{V}_{h+1}^+ + H.\end{aligned} \tag{184}$$

For any $h \in [H]$, there exists at least one state $s_h^\star$ that satisfies $\widehat{V}_h^+(s_h^\star) = \min_{s\in\mathcal{S}} \widehat{V}_h^+(s)$. Furthermore, for any accessible uncertainty set $\sigma^+ > 0$ and $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, we define an auxiliary vector $\widehat{P}'_{h,s,a,b} \in \mathbb{R}^S$ by reducing the values of several elements of $\widehat{P}^0_{h,s,a,b}$ strictly, namely,

$$0 \le \widehat{P}'_{h,s,a,b} \le \widehat{P}^0_{h,s,a,b} \quad \text{and} \quad \sum_{s'\in\mathcal{S}} \widehat{P}^0_{h,s,a,b}(s') - \widehat{P}'_{h,s,a,b}(s') = \left\|\widehat{P}'_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\right\|_1 = \sigma^+. \tag{185}$$

Let $l_{s_h^\star}$ represent an S-dimensional standard basis under $s_h^\star$. We can derive that

$$\frac{1}{2}\left\|\widehat{P}'_{h,s,a,b} + \sigma^+\left[l_{s_h^\star}\right]^\top - \widehat{P}^0_{h,s,a,b}\right\|_1 \le \frac{1}{2}\left\|\widehat{P}'_{h,s,a,b} - \widehat{P}^0_{h,s,a,b}\right\|_1 + \frac{1}{2}\left\|\sigma^+\left[l_{s_h^\star}\right]^\top\right\|_1 \le \sigma^+, \tag{186}$$

where the first inequality is valid since the 'distance' function (e.g., TV distance) satisfies the triangle inequality.

Therefore, we can conclude that $\widehat{P}'_{h,s,a,b} + \sigma^+\left[l_{s_h^\star}\right]^\top \in \mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})$ and $\widehat{P}'_{h,s,a,b} + \sigma^+\left[l_{s_h^\star}\right]^\top$ is a distribution vector based on (186), leading to

$$\begin{aligned}
\inf_{P\in\mathcal{U}^{\sigma^+}(\widehat{P}^0_{h,s,a,b})} P\widehat{V}^+_{h+1} &\le \left(\widehat{P}'_{h,s,a,b} + \sigma^+\left[l_{s^\star_{i,h}}\right]^\top\right)\widehat{V}^+_{h+1} \\
&\le \left\|\widehat{P}'_{h,s,a,b}\right\|_1 \left\|\widehat{V}^+_{h+1}\right\|_\infty + \sigma^+\widehat{V}^+_{h+1}(s^\star_{h+1}) \\
&\le \left(1-\sigma^+\right)\max_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s) + \sigma^+\min_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s),
\end{aligned} \tag{187}$$

where the last inequality holds since

$$\left\|P'_{h,s,a,b}\right\|_1 = \sum_{s'} P'_{h,s,a,b}(s') = -\sum_{s'}\left(P^0_{h,s,a,b}(s') - P'_{h,s,a,b}(s')\right) + \sum_{s'} P^0_{h,s,a,b}(s') = 1-\sigma^+. \tag{188}$$

Putting (187) and (184) together shows

$$\begin{aligned}
\max_{s\in\mathcal{S}}\widehat{V}^+_h(s) &\le 1 + \max_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} \inf_{P\in\mathcal{U}^{\sigma^+}(P^0_{h,s,a,b})} P\widehat{V}^+_{h+1} + H \\
&\le H + 1 + \left(1-\sigma^+\right)\max_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s) + \sigma^+\min_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s).
\end{aligned} \tag{189}$$

By substituting (189) it (183), it readily follows that

$$\begin{aligned}
\max_{s\in\mathcal{S}}\widehat{V}^+_h - \min_{s\in\mathcal{S}}\widehat{V}^+_h \le& H + 1 + \left(1-\sigma^+\right)\max_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s) + \sigma^+\min_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s) - \min_{s\in\mathcal{S}}V^+_{h+1}(s) \\
=& H + 1 + (1-\sigma^+)\left(\max_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s) - \min_{s\in\mathcal{S}}\widehat{V}^+_{h+1}(s)\right) \\
\le& H + 1 + (1-\sigma^+)\left[H + 1 + (1-\sigma^+)\left(\max_{s\in\mathcal{S}}\widehat{V}^+_{h+2}(s) - \min_{s\in\mathcal{S}}\widehat{V}^+_{h+2}(s)\right)\right] \\
\le& \cdots \le \frac{(H+1)\left(1-(1-\sigma^+)^{H-h}\right)}{\sigma^+}.
\end{aligned} \tag{190}$$

which, combined with $\max_{s\in\mathcal{S}}\widehat{V}^+_h(s) - \min_{s\in\mathcal{S}}\widehat{V}^+_h(s) \le H$, conclude this proof.

G.4 Proof of Lemma D.4

First, we introduce auxiliary values and reward functions to control $\sum_{i=1}^H \sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^0_{i,s,\mu^{\mathsf{d}}(s),b}}\left(\widehat{V}\right)$ as below: for any time step i

- $\widehat{V}_i^{\mathsf{m}} \coloneqq \min_{s\in\mathcal{S}}\widehat{V}_i^+(s)$: the minimum value of all the entries in vector $\widehat{V}_i^+$.
- $\widehat{V}_i' \coloneqq \widehat{V}_i^+ - \widehat{V}_i^{\mathsf{m}}1$: truncated value function.
- $\widehat{r}_i^{\mu^{\mathsf{d}},\nu^\star}(s) = \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))}\widehat{r}_i(s,a,b)$: average reward function.

- $\widehat{r}_i^{\mathsf{m}} = r_i^{\mu^{\mathsf{d}},\nu^\star} + \left(\widehat{V}_{i+1}^{\mathsf{m}} - \widehat{V}_i^{\mathsf{m}}\right) 1$: truncated reward function.

Applying the robust Bellman's consistency equation in (30) gives

$$\begin{aligned}
\widehat{V}_i' = \widehat{V}_i^+ - \widehat{V}_i^{\mathsf{m}} 1 &\overset{\text{(i)}}{\leq} \widehat{r}_i^{\mu^{\mathsf{d}},\nu^\star} + \widetilde{P}_i^{\inf,\widehat{V}} \widehat{V}_{i+1}^+ + 2\beta_i^{\mu^{\mathsf{d}},\nu^\star} - \widehat{V}_i^{\mathsf{m}} 1 \\
&= \widehat{r}_i^{\mu^{\mathsf{d}},\nu^\star} + \widetilde{P}_i^{\inf,\widehat{V}} \widehat{V}_{i+1}^+ + \left(\widehat{V}_{i+1}^{\mathsf{m}} 1 - \widehat{V}_i^{\mathsf{m}} 1\right) - \widehat{V}_{i+1}^{\mathsf{m}} 1 + 2\beta_i^{\mu^{\mathsf{d}},\nu^\star} \\
&= \widehat{r}_i^{\mathsf{m}} + \widetilde{P}_i^{\inf,\widehat{V}} \widehat{V}_{i+1}^+ - \widehat{V}_{i+1}^{\mathsf{m}} 1 + 2\beta_i^{\mu^{\mathsf{d}},\nu^\star} \\
&= \widehat{r}_i^{\mathsf{m}} + \widetilde{P}_i^{\inf,\widehat{V}} \widehat{V}_{i+1}' + 2\beta_i^{\mu^{\mathsf{d}},\nu^\star},
\end{aligned} \tag{191}$$

where (i) follows from the fact that

$$\begin{aligned}
\widehat{V}_i^+(s) \leq& \widehat{r}_i^{\mu^{\mathsf{d}},\nu^\star}(s) + \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{i,s,a,b}\right)} P\widehat{V}_{i+1}^+ + \beta_i^{\mu^{\mathsf{d}},\nu^\star}(s) \\
\overset{\text{(i)}}{\leq}& \widehat{r}_i^{\mu^{\mathsf{d}},\nu^\star}(s) + \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \Bigg[\inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{i,s,a,b}\right)} P\widehat{V}_{i+1}^+ \\
&+ \left| \inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{i,s,a,b}\right)} P\widehat{V}_{i+1}^+ - \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{i,s,a,b}\right)} P\widehat{V}_{i+1}^+ \right| \Bigg] + \beta_i^{\mu^{\mathsf{d}},\nu^\star}(s) \\
\overset{\text{(ii)}}{\leq}& \widehat{r}_i^{\mu^{\mathsf{d}},\nu^\star}(s) + \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \left[P_{i,s,a,b}^{\inf,\widehat{V}} \widehat{V}_{i+1}^+ \right] + 2\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s) \\
\overset{\text{(iii)}}{=}& \widehat{r}_i^{\mu^{\mathsf{d}},\nu^\star}(s) + \widetilde{P}_{i,s}^{\inf,\widehat{V}} \widehat{V}_{i+1}^+ + 2\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s),
\end{aligned} \tag{192}$$

(ii) is valid under the notation

$$P_{i,s,a,b}^{\inf,\widehat{V}} \coloneqq \operatorname{argmin}_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{i,s,a,b}\right)} P\widehat{V}_{i+1}^+, \tag{193}$$

and (iii) holds under the notation as $\widetilde{P}_{i,s}^{\inf,\widehat{V}} \coloneqq \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} P_{i,s,a,b}^{\inf,\widehat{V}}$ and the sequence as $\widetilde{P}_i^{\inf,V} \in \mathbb{R}^{S\times S}$. Besides, (i) in (192) exists due to (48) in Lemma D.1 for $N_i(s,a,b) > 0$ and

$$\left| \inf_{P\in\mathcal{U}^{\sigma^+}\left(P^0_{i,s,a,b}\right)} P\widehat{V}_{i+1}^+ - \inf_{P\in\mathcal{U}^{\sigma^+}\left(\widehat{P}^0_{i,s,a,b}\right)} P\widehat{V}_{i+1}^+ \right| \leq H = \beta_i^{\mu^{\mathsf{d}},\nu^\star}(s) \text{ for } N_i(s,a,b) = 0. \tag{194}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \mathsf{Var}_{P_{i,s,a,b}^{\inf,V}}\left(\widehat{V}_{i+1}^+\right) \\
\overset{\text{(i)}}{=}& \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \mathsf{Var}_{P_{i,s,a,b}^{\inf,V}}\left(\widehat{V}_{i+1}'\right) \\
=& \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \left[P_{i,s,a,b}^{\inf,V}\left(\widehat{V}_{i+1}' \circ \widehat{V}_{i+1}'\right) - \left(P_{i,s,a,b}^{\inf,V}\widehat{V}_{i+1}'\right) \circ \left(P_{i,s,a,b}^{\inf,V}\widehat{V}_{i+1}'\right) \right] \\
\leq& \mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))} \left[P_{i,s,a,b}^{\inf,V}\left(\widehat{V}_{i+1}' \circ \widehat{V}_{i+1}'\right) - \left(P_{i,s,a,b}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right) \circ \left(P_{i,s,a,b}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right) \right] \\
\overset{\text{(ii)}}{\leq}& \widetilde{P}_{i,s}^{\inf,V}\left(\widehat{V}_{i+1}' \circ \widehat{V}_{i+1}'\right) - \left(\widetilde{P}_{i,s}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right) \circ \left(\widetilde{P}_{i,s}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right) \\
=& \widetilde{P}_{i,s}^{\inf,V}\left(\widehat{V}_{i+1}' \circ \widehat{V}_{i+1}'\right) - \widehat{V}_i'(s) \circ \widehat{V}_i'(s) + \widehat{V}_i'(s) \circ \widehat{V}_i'(s) - \left(\widetilde{P}_{i,s}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right) \circ \left(\widetilde{P}_{i,s}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right) \\
=& \widetilde{P}_{i,s}^{\inf,V}\left(\widehat{V}_{i+1}' \circ \widehat{V}_{i+1}'\right) - \widehat{V}_i'(s) \circ \widehat{V}_i'(s) + \left(\widehat{V}_i'(s) - \left(\widetilde{P}_{i,s}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right)\right) \circ \left(\widehat{V}_i'(s) + \left(\widetilde{P}_{i,s}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right)\right) \\
\overset{\text{(iii)}}{\leq}& \widetilde{P}_{i,s}^{\inf,V}\left(\widehat{V}_{i+1}' \circ \widehat{V}_{i+1}'\right) - \widehat{V}_i'(s) \circ \widehat{V}_i'(s) + \left(\widehat{r}_i^{\mathsf{m}}(s) + 2\beta_h^{\mu^{\mathsf{d}},\nu^\star}(s)\right) \circ \left(\widehat{V}_i'(s) + \left(\widetilde{P}_{i,s}^{\inf,\widehat{V}}\widehat{V}_{i+1}'\right)\right) \\
\overset{\text{(iv)}}{\leq}& \widetilde{P}_{i,s}^{\inf,V}\left(\widehat{V}_{i+1}' \circ \widehat{V}_{i+1}'\right) - \widehat{V}_i'(s) \circ \widehat{V}_i'(s) + \left(\left\|\widehat{V}_i'\right\|_\infty + \left\|\widehat{V}_{i+1}'\right\|_\infty\right)\left(2\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s) + 1\right),
\end{aligned} \tag{195}$$

where (i) follows from that $\mathrm{Var}_{P_{i,s}^{\inf,V}}(V - b1) = \mathrm{Var}_{P_{i,s}^{\inf,V}}(V)$ for any value vector $V \in \mathbb{R}^S$ and scalar b, (ii) holds since

$$\begin{aligned}
&\mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))}\left[\left(P_{i,s,a,b}^{\inf,\widehat{V}}\widehat{V}'_{i+1}\right)\circ\left(P_{i,s,a,b}^{\inf,\widehat{V}}\widehat{V}'_{i+1}\right)\right]\\
\geq&\mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))}\left[\left(P_{i,s,a,b}^{\inf,\widehat{V}}\widehat{V}'_{i+1}\right)\right]\circ\mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))}\left[\left(P_{i,s,a,b}^{\inf,\widehat{V}}\widehat{V}'_{i+1}\right)\right],
\end{aligned}$$

(iii) comes from (191), and (iv) arises from $\widehat{r}_i^{\mathsf{m}} \leq r_i \leq 1$ due to $\widehat{V}_{i+1}^{\mathsf{m}} - \widehat{V}_i^{\mathsf{m}} \leq 0$ by definition. Consequently, combining (62), we arrive at

$$\begin{aligned}
&\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P_{i,s,a,b}^{\inf,V}}\left(\widehat{V}_{i+1}^{+}\right)\\
=&\sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\mathbb{E}_{(a,b)\sim(\mu^{\mathsf{d}}(s),\nu^\star(s))}\mathsf{Var}_{P_{i,s,a,b}^{\inf,V}}\left(\widehat{V}_{i+1}^{+}\right)\\
\leq&\sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\Big(\widetilde{P}_{i,s}^{\inf,V}\left(\widehat{V}'_{i+1}\circ\widehat{V}'_{i+1}\right)-\widehat{V}'_i(s)\circ\widehat{V}'_i(s)\\
&+\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\left(2\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s)+1\right)\Big)\\
\leq&\sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\left(\widetilde{P}_{i,s}^{\inf,V}\left(\widehat{V}'_{i+1}\circ\widehat{V}'_{i+1}\right)-\widehat{V}'_i(s)\circ\widehat{V}'_i(s)\right)+\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\\
&+2\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s)\\
=&\sum_{s\in\mathcal{S}}\left(d_{i+1}^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\left(\widehat{V}'_{i+1}(s)\circ\widehat{V}'_{i+1}(s)\right)-d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\widehat{V}'_i(s)\circ\widehat{V}'_i(s)\right)+\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\\
&+2\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s)\\
=&\sum_{s\in\mathcal{S}}\left(d_{i+1}^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\left(\widehat{V}'_{i+1}(s)\circ\widehat{V}'_{i+1}(s)\right)-d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\widehat{V}'_i(s)\circ\widehat{V}'_i(s)\right)+\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\\
&+2\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V}). \qquad (196)
\end{aligned}$$

Besides, under TV distance, we have

$$\begin{aligned}
\left|\mathsf{Var}_{P_{i,s,a,b}^{0}}\left(\widehat{V}_{i+1}^{+}\right)-\mathsf{Var}_{P_{i,s,a,b}^{\inf,V}}\left(\widehat{V}_{i+1}^{+}\right)\right|&=\left|\mathsf{Var}_{P_{i,s,a,b}^{0}}\left(\widehat{V}'_{i+1}\right)-\mathsf{Var}_{P_{i,s,a,b}^{\inf,V}}\left(\widehat{V}'_{i+1}\right)\right|\\
&\leq\left\|P_{i,s,a,b}^{0}-P_{i,s,a,b}^{\inf,V}\right\|_1\left\|\widehat{V}'_{i+1}\right\|_\infty^2\\
&\leq\sigma^+\left\|\widehat{V}'_{i+1}\right\|_\infty^2\leq(H+1)\left\|\widehat{V}'_{i+1}\right\|_\infty, \qquad (197)
\end{aligned}$$

where the last inequality comes from Lemma D.3.

Therefore, we derive

$$
\begin{aligned}
&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^0_{i,s,a,b}}\big(\widehat{V}^+_{i+1}\big)\\
\leq&\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\mathsf{Var}_{P^{\inf,V}_{i,s,a,b}}\big(\widehat{V}^+_{i+1}\big)\\
&+\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\left|\mathsf{Var}_{P^0_{i,s,a,b}}\big(\widehat{V}^+_{i+1}\big)-\mathsf{Var}_{P^{\inf,V}_{i,s,a,b}}\big(\widehat{V}^+_{i+1}\big)\right|\\
\leq&\sum_{i=1}^{H}2\left(\left\|\widehat{V}'_i\right\|_\infty+\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\sum_{s\in\mathcal{S}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\beta_i^{\mu^{\mathsf{d}},\nu^\star}(s)+\sum_{i=1}^{H}\left(\left\|\widehat{V}'_i\right\|_\infty+(H+2)\left\|\widehat{V}'_{i+1}\right\|_\infty\right)\\
&+\sum_{s\in\mathcal{S}} d_{H+1}^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s)\widehat{V}'_{H+1}(s)\circ\widehat{V}'_{H+1}(s)\\
\leq&4\sum_{i=1}^{H}\min\left\{\frac{(H+1)\left(1-(1-\sigma^+)^{H-i}\right)}{\sigma^+},H\right\}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})\\
&+(H+3)\sum_{i=1}^{H}\min\left\{\frac{(H+1)\left(1-(1-\sigma^+)^{H-i}\right)}{\sigma^+},H\right\}\\
\overset{\text{(i)}}{\leq}&4\sum_{i=1}^{H}\min\left\{\frac{(H+1)\left(1-(1-\sigma^+)^{H-i}\right)}{\sigma^+},H\right\}\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})\\
&+(H+3)\sum_{i=1}^{H}\min\left\{\frac{(H+1)\left(1-(1-\sigma^+)^{H-i}\right)}{\sigma^+},H\right\}\\
\overset{\text{(ii)}}{\leq}&4H\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\}\sum_{i=1}^{H}\sum_{(s,b)\in\mathcal{S}\times\mathcal{B}} d_i^{\mathsf{p},\mu^{\mathsf{d}},\nu^\star}(s,\mu^{\mathsf{d}}(s),b)\beta_i(s,\mu^{\mathsf{d}}(s),b,\widehat{V})\\
&+(H+3)H\min\left\{\frac{2(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2},H\right\},
\end{aligned}
\tag{198}
$$

where (i) comes from Cauchy-Schwarz inequality and the (ii) holds since

$$
\begin{aligned}
\sum_{i=1}^{H}\frac{(H+1)\left(1-(1-\sigma^+)^{H-i}\right)}{\sigma^+}=&\frac{H(H+1)}{\sigma^+}-\sum_{i=0}^{H-1}\frac{(H+1)(1-\sigma^+)^i}{\sigma^+}\\
=&\frac{H(H+1)}{\sigma^+}-\frac{(H+1)(1-(1-\sigma^+)^H)}{(\sigma^+)^2}\\
=&\frac{(H+1)(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2}\\
\leq&\frac{2H(H\sigma^+-1+(1-\sigma^+)^H)}{(\sigma^+)^2}.
\end{aligned}
$$

H Proof for Theorem 4.4

To establish Theorem 4.4, we present key auxiliary facts, each addressing critical components of the proof and providing essential theoretical underpinnings.

H.1 Proof of Lemma E.2

Since all RMDPs in $\mathcal{M}(\mathcal{F},\Phi)$ are constructed similarly for each $w\in\mathcal{F}$ and $\phi\in\Phi$, we will focus on a specific RMDP $\mathcal{M}_f^\phi\in\mathcal{M}(\mathcal{F},\Phi)$, with the results applicable to all other RMDPs in $\mathcal{M}(\mathcal{F},\Phi)$.

H.1.1 Ordering the robust value function over different states

Before proceeding, we introduce several facts and notations that will be useful throughout this section. First, for any $\mathcal{M}_f^\phi$ and any policy $\widetilde{\mu}$, we observe the following at the final step $H+1$:

$$\forall s \in \mathcal{M} \cup \mathcal{N}: \quad V_{H+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s) = 0. \tag{199}$$

Then for step H, we can easily verify that

$$\forall s \in \mathcal{N}: \quad V_H^{\widetilde{\mu},\sigma^+,f,\phi}(s) = \mathbb{E}_{a\sim\widetilde{\mu}_H(\cdot\,|\,s)}\left[r_H(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^{\sigma^+}(P_{H,s,a}^{f,\phi})} \mathcal{P}V_{H+1}^{\widetilde{\mu},\sigma^+,f,\phi}\right] = 1 \tag{200a}$$

$$\forall s \in \mathcal{M}: \quad V_H^{\widetilde{\mu},\sigma^+,f,\phi}(s) = \mathbb{E}_{a\sim\widetilde{\mu}_H(\cdot\,|\,s)}\left[r_H(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^{\sigma^+}(P_{H,s,a}^{f,\phi})} \mathcal{P}V_{H+1}^{\widetilde{\mu},\sigma^+,f,\phi}\right] = 0, \tag{200b}$$

which holds by (199) and the definition of the reward function (see (87)). The above fact directly indicates that

$$\forall (s,s') \in \mathcal{M}\times\mathcal{N}: \quad \min_{\widetilde{s}\in\mathcal{S}} V_H^{\widetilde{\mu},\sigma^+,f,\phi}(\widetilde{s}) = V_H^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \le V_H^{\widetilde{\mu},\sigma^+,f,\phi}(s) < V_H^{\widetilde{\mu},\sigma^+,f,\phi}(s'), \tag{201a}$$

$$\forall (s,s') \in \mathcal{N}\times\mathcal{N}: \quad V_H^{\widetilde{\mu},\sigma^+,f,\phi}(s) = V_H^{\widetilde{\mu},\sigma^+,f,\phi}(s'). \tag{201b}$$

Then we introduce a claim which we will prove by induction in a moment as below:

$$\forall (h,s,s') \in [H]\times\mathcal{M}\times\mathcal{N}: \quad V_h^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \le V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s) < V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s') \tag{202a}$$

$$\forall (s,s') \in \mathcal{N}\times\mathcal{N}: \quad V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s) = V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s'). \tag{202b}$$

Note that the base case when the time step is $H+1$ is verified in (201). Assume that the following fact at time step $h+1$ holds

$$\forall (s,s') \in \mathcal{M}\times\mathcal{N}: \quad \min_{\widetilde{s}\in\mathcal{S}} V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(\widetilde{s}) = V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \le V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s) < V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s'), \tag{203a}$$

$$\forall (s,s') \in \mathcal{N}\times\mathcal{N}: \quad V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s) = V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s'). \tag{203b}$$

The rest of the proof focuses on proving the same property for time step h. For RMDP $\mathcal{M}_f^\phi \in \mathcal{M}(\mathcal{F},\Phi)$ and any policy $\widetilde{\mu}$, we characterize the robust value function of different states separately:

- *For state $s \in \mathcal{N}$:* we observe that for any $s \in \mathcal{N}$,

$$\begin{aligned} V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s) &= \mathbb{E}_{a\sim\widetilde{\mu}_h(\cdot\,|\,s)}\left[r_h(s,a) + \inf_{P\in\mathcal{U}^{\sigma^+}(P_{h,s,a}^{f,\phi})} PV_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}\right] \\ &\overset{\text{(i)}}{=} 1 + \mathbb{E}_{a\sim\widetilde{\mu}_h(\cdot\,|\,s)}\left[P_h^{\inf,f,\phi}(s\,|\,s,a)V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s)\right] + \sigma^+ V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \\ &= 1 + (1-\sigma^+)V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(s) + \sigma^+ V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f), \end{aligned} \tag{204}$$

  where (i) holds by $r_h(s,a) = 1$ for all $s \in \mathcal{N}$ (see (87)), the fact that $\min_{\widetilde{s}\in\mathcal{S}} V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(\widetilde{s}) = V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)$ induced by the induction assumption (cf. (203)) and the definition of $P_h^{\inf,f,\phi}(s\,|\,s,a)$ in (90), and the last equality follows from $P^{f,\phi}(s\,|\,s,a) = 1$ for all $(s,a) \in \mathcal{N}\times\mathcal{A}_{\mathsf{one}}$. Resorting to the induction assumption in (203), we have

$$\forall (s,s') \in \mathcal{N}\times\mathcal{N}: \quad V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s) = V_h^{\widetilde{\mu},\sigma^+,f,\phi}(s'). \tag{205}$$

- *For state m_f:* first, the robust value function at state m_f obeys

$$\begin{aligned}
V_h^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) &= \mathbb{E}_{a\sim\widetilde{\mu}_h(\cdot\,|\,m_f)}\left[r_h(m_f,a) + \inf_{P\in\mathcal{U}^{\sigma^+}(P^{f,\phi}_{h,m_f,a})} PV_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}\right] \\
&\overset{\text{(i)}}{=} 0 + \widetilde{\mu}_h(\phi_h\,|\,m_f)\inf_{P\in\mathcal{U}^{\sigma^+}(P^{f,\phi}_{h,m_f,\phi_h})} PV_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi} \\
&\quad + \widetilde{\mu}_h(1-\phi_h\,|\,m_f)\inf_{P\in\mathcal{U}^{\sigma^+}(P^{f,\phi}_{h,m_f,1-\phi_h})} PV_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi} \\
&\overset{\text{(ii)}}{=} \widetilde{\mu}_h(\phi_h\,|\,m_f)\Big[p^{\inf}V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + \big(1-p^{\inf}\big)\,V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\Big] \\
&\quad + \widetilde{\mu}_h(1-\phi_h\,|\,m_f)\Big[q^{\inf}V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + \big(1-q^{\inf}\big)\,V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\Big] \\
&\overset{\text{(iii)}}{=} m_h^{\widetilde{\mu},f,\phi}V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + (1-m_h^{\widetilde{\mu},f,\phi})V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \qquad (206)\\
&\le (1-\sigma^+)V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + \sigma^+V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f). \qquad (207)
\end{aligned}$$

  where (i) uses the definition of the robust value function and the reward function in (87), (ii) uses the induction assumption in (203) so that the minimum is attained by picking the choice specified in (91) to absorb probability mass to state m_f, and (iii) holds by plugging in the definition (93) of $m_h^{\widetilde{\mu},f,\phi}$. Finally, the last inequality follows from the fact that function $f(m) \coloneqq mV_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + (1-m)V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)$ is monotonically increasing with m since $V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) > V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)$ (see the induction assumption (203)), and the fact $m_h^{\widetilde{\mu},f,\phi} \le 1-\sigma^+$.

Combining the above results with (205), we confirm the claim in (202).

H.1.2 Deriving the optimal policy and optimal robust value function

We shall characterize the optimal policy and corresponding optimal robust value function for different states separately:

- *For states in $\mathcal{M}$:* Recall (206)

$$V_h^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) = m_h^{\widetilde{\mu},f,\phi}V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + (1-m_h^{\widetilde{\mu},f,\phi})V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \qquad (208)$$

  and the fact $V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) > V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)$ in (202). We observe that (208) is monotonicity increasing w.r.t. $m_h^{\widetilde{\mu},f,\phi}$, and $m_h^{\widetilde{\mu},f,\phi}$ is also increasing in $\widetilde{\mu}_h(\phi_h\,|\,m_f)$ (refer to the fact $p^{\inf} \ge q^{\inf}$ since $p \ge q$; see (99) and (91)). Consequently, the optimal policy and optimal robust value function in state m_f thus obey

$$\begin{aligned}
\forall h\in[H]:\quad & \widetilde{\mu}_h^{\star,f,\phi}(\phi_h\,|\,m_f) = 1, \\
& V_h^{\star,\sigma^+,f,\phi}(m_f) = p^{\inf}V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) + \big(1-p^{\inf}\big)\,V_{h+1}^{\star,\sigma^+,f,\phi}(m_f). \qquad (209)
\end{aligned}$$

- *For states $s\in\mathcal{N}$:* Recall the transitions in (98) and (86). Considering that the action does not influence the state transition for all states $s\in\mathcal{N}$, without loss of generality, we choose the robust optimal policy obeying

$$\forall s\in\mathcal{N}:\quad \widetilde{\mu}_h^{\star,f,\phi}(\phi_h\,|\,s) = 1. \qquad (210)$$

H.2 Proof of (105)

With the initial state distribution and behavior policy defined in (97), we have for any MDP $\mathcal{M}_\phi^f$,

$$d_1^{\mathsf{n},P^{\phi,f}}(s) = \varrho^{\mathsf{n}}(s) = \overline{\varrho}(s),$$

which leads to

$$\forall (m_f, a) \in \mathcal{M} \times \mathcal{A}_{\mathsf{one}}: \quad d_1^{\mathsf{n},P^{\phi},f}(m_f, a) = \frac{1}{2}\overline{\varrho}(m_f). \tag{211}$$

Along with $d_1^{\mathsf{n},P^{\phi},f}(n_f, a) = \frac{1}{2}\overline{\varrho}(n_f) = 0$, (105a) is proved.

In view of (98), the state occupancy distribution at any step $h = 2, 3, \cdots, H$ obeys

$$\begin{aligned} d_h^{\mathsf{n},P^{\phi},f}(m_f) &\geq \mathbb{P}\left\{s_h = s' \,|\, s_{h-1} = m_f; \widetilde{\mu}^{\mathsf{n}}\right\} \\ &\geq d_{h-1}^{\mathsf{n},P^{\phi},f}(m_f)\left[\widetilde{\mu}_{h-1}^{\mathsf{n}}(\phi_{h-1} \,|\, m_f)(1-p-\Delta) + \widetilde{\mu}_{h-1}^{\mathsf{n}}(1-\phi_{h-1} \,|\, m_f)(1-p)\right] \\ &\geq d_{h-1}^{\mathsf{n},P^{\phi},f}(m_f)(1-p-\Delta) \geq \cdots \geq d_1^{\mathsf{n},P^{\phi}}(m_f)\prod_{j=0}^{h-1}(1-p-\Delta) \\ &\geq d_1^{\mathsf{n},P^{\phi}}(m_f)\Big(1-p-\Delta\Big)^H > \frac{\overline{\varrho}(m_f)}{2}, \end{aligned} \tag{212}$$

where the last line makes use of the properties p and Δ in (101) and

$$\Big(1-p-\Delta\Big)^H \geq \Big(1-\frac{c_2}{2H}\Big)^H \geq \Big(1-\frac{1}{2H}\Big)^H \geq \frac{1}{2},$$

provided that $0 < c_2 < 1$. In addition, as state n_f is an absorbing state and state m_f will only transfer to itself or state n_f at each time step, we directly achieve that

$$d_h^{\mathsf{n},P^{\phi},f}(m_f) \leq d_{h-1}^{\mathsf{n},P^{\phi},f}(m_f) \leq \cdots \leq d_1^{\mathsf{n},P^{\phi},f}(m_f) \leq \overline{\varrho}(m_f). \tag{213}$$

For state n_f, as it is absorbing, we directly have

$$d_h^{\mathsf{n},P^{\phi},f}(n_f) = \mathbb{P}\left\{s_h = n_f \,|\, s_{h-1} = n_f; \widetilde{\mu}^{\mathsf{n}}\right\} \geq d_{h-1}^{\mathsf{n},P^{\phi},f}(n_f) \geq \cdots \geq d_1^{\mathsf{n},P^{\phi},f}(n_f) = \overline{\varrho}(n_f). \tag{214}$$

According to the assumption in (104), it is easily verified that

$$d_h^{\mathsf{n},P^{\phi},f}(n_f) \leq 1 \leq 2\overline{\varrho}(n_f). \tag{215}$$

Finally, combining (212), (213), (214), (215), the definitions of $P_h^{\star}(\cdot \,|\, s, a)$ in (98), and the Markov property, we arrive at for any $(h, s) \in [H] \times \mathcal{S}$,

$$\frac{\overline{\varrho}(s)}{2} \leq d_h^{\mathsf{n},P^{\phi},f}(s) \leq 2\overline{\varrho}(s), \tag{216}$$

which directly leads to

$$\frac{\overline{\varrho}(s)}{4} \leq d_h^{\mathsf{n},P^{\phi},f}(s, a) = \widetilde{\mu}_1^{\mathsf{n}}(a \,|\, s) d_h^{\mathsf{n},P^{\phi},f}(s) \leq \overline{\varrho}(s). \tag{217}$$

H.3 Proof of (107)

Examining the definition of $C_{\mathrm{r}}^{\star}$ in (20), we make the following observations.

- For $h = 1$, we have

$$\begin{aligned} &\max_{(s,a,P)\in\mathcal{S}_{\mathsf{one}}\times\mathcal{A}_{\mathsf{one}}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_1^{\star,P}(s,a), \frac{1}{4SA}\right\}}{d_1^{\mathsf{n},P^{\phi},f}(s,a)} \\ &\overset{\text{(i)}}{=} \max_{(s,P)\in\mathcal{M}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_1^{\star,P}(s,\phi_1), \frac{1}{4SA}\right\}}{d_1^{\mathsf{n},P^{\phi},f}(s,\phi_1)} \\ &\overset{\text{(ii)}}{=} \max_{(s,P)\in\mathcal{M}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{1}{4SAd_1^{\mathsf{n},P^{\phi},f}(s,\phi_1)} \\ &\overset{\text{(iii)}}{=} \max_{s\in\mathcal{M}} \frac{1}{2SA\overline{\varrho}(s)} = C, \end{aligned} \tag{218}$$

  where (i) holds by $d_1^{\star,P}(s) = \rho(s) = 0$ for all $s \in \mathcal{N}$ (see (106)) and $\widetilde{\mu}_h^{\star,\phi}(\phi_h \,|\, s) = 1$ for all $(s, h) \in \mathcal{M} \times [H]$ (see (95)), (ii) follows from the fact that $d_1^{\star,P}(s, \phi_1) = 1$ for all $s \in \mathcal{M}$, (iii) is verified in (105), and the last equality arises from the definition in (103).

- Similarly, for $h = 2, 3, \cdots, H$, we arrive at

$$\begin{aligned}
&\max_{(s,a,P)\in\mathcal{S}_{\mathsf{one}}\times\mathcal{A}_{\mathsf{one}}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{4SA}\right\}}{d_h^{\mathsf{n},P^{\phi,f}}(s,a)} \\
&\overset{\text{(i)}}{=} \max_{(s,P)\in\mathcal{S}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_h^{\star,P}(s,\phi_h), \frac{1}{4SA}\right\}}{d_h^{\mathsf{n},P^{\phi,f}}(s,\phi_h)} \\
&\leq \max_{(s,P)\in\mathcal{M}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{1}{4SAd_h^{\mathsf{n},P^{\phi,f}}(s,\phi_h)} \\
&\overset{\text{(ii)}}{\leq} \max_{s\in\mathcal{M}} \frac{1}{2SA\overline{\varrho}(s)} = 2C,
\end{aligned} \tag{219}$$

  where (i) holds by the optimal policy in (95) and the trivial fact that $d_h^{\star,P}(s) = 0$ for all $s \in \mathcal{N}$ (see (106) and (98)), (ii) arises from (105), and the last equality comes from (103).

Combining the above cases, we complete the proof by

$$\frac{C}{2} \leq C_{\mathrm{r}}^{\star} = \max_{(h,s,a,P)\in[H]\times\mathcal{S}_{\mathsf{one}}\times\mathcal{A}_{\mathsf{one}}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{4SA}\right\}}{d_h^{\mathsf{n},P^{\phi,f}}(s,a)} \leq C.$$

H.4 Proof of (112)

Recalling (94a) and (96), we first consider a more general form

$$\begin{aligned}
&V_h^{\star,\sigma^+,f,\phi}(m_f) - V_h^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \\
=&p^{\inf}V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) + (1-p^{\inf})V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) \\
&- \left(m_h^{\widetilde{\mu},f,\phi}V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f) + \left[1-m_h^{\widetilde{\mu},f,\phi}\right]V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right) \\
=&\left(p^{\inf}-m_h^{\widetilde{\mu},f,\phi}\right)V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) + m_h^{\widetilde{\mu},f,\phi}\left(V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f)\right) \\
&+ (1-p^{\inf})\left(V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right) - \left(p^{\inf}-m_h^{\widetilde{\mu},f,\phi}\right)V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \\
=&m_h^{\widetilde{\mu},f,\phi}\left(V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(n_f)\right) + (1-p^{\inf})\left(V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right) \\
&+ \left(p^{\inf}-m_h^{\widetilde{\mu},f,\phi}\right)\left(V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) - V_{h+1}^{\star,\sigma^+,f,\phi}(m_f)\right) \\
\geq&(1-p^{\inf})\left(V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right) \\
&+ \left(p^{\inf}-m_h^{\widetilde{\mu},f,\phi}\right)\left(V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) - V_{h+1}^{\star,\sigma^+,f,\phi}(m_f)\right) \\
\geq&(1-p^{\inf})\left(V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right) \\
&+ \frac{1}{2}(p-q)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f) - \widetilde{\mu}_h(\cdot\,|\,m_f)\right\|_1\left(V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) - V_{h+1}^{\star,\sigma^+,f,\phi}(m_f)\right),
\end{aligned} \tag{220}$$

where the last inequality holds since

$$\begin{aligned}
p^{\inf}-m_h^{\widetilde{\mu},f,\phi} &= \left(p^{\inf}-q^{\inf}\right)\left(1-\widetilde{\mu}_h(\phi_h\,|\,m_f)\right) \\
&= (p-q)\left(1-\widetilde{\mu}_h(\phi_h\,|\,m_f)\right) \\
&= \frac{1}{2}(p-q)\left(1-\widetilde{\mu}_h(\phi_h\,|\,m_f) + \widetilde{\mu}_h(1-\phi_h\,|\,m_f)\right) \\
&= \frac{1}{2}(p-q)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f) - \widetilde{\mu}_h(\cdot\,|\,m_f)\right\|_1,
\end{aligned} \tag{221}$$

with the first equality holding by (93) and the second existing by (91).

To further control (220), we have

$$\begin{aligned}
V_h^{\star,\sigma^+,f,\phi}(n_f) - V_h^{\star,\sigma^+,f,\phi}(m_f) \overset{\text{(i)}}{=} & 1 + (1-\sigma^+)V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) + \sigma^+ V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) \\
& - \left(p^{\inf} V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) + (1-p^{\inf})V_{h+1}^{\star,\sigma^+,f,\phi}(m_f)\right) \\
= & 1 + (1-p^{\inf}-\sigma^+)\left(V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) - V_{h+1}^{\star,\sigma^+,f,\phi}(m_f)\right) \\
\overset{\text{(ii)}}{=} & 1 + (1-p)\left(V_{h+1}^{\star,\sigma^+,f,\phi}(n_f) - V_{h+1}^{\star,\sigma^+,f,\phi}(m_f)\right) \\
= & \cdots = \sum_{j=0}^{H-h}(1-p)^j, \qquad (222)
\end{aligned}$$

where (i) follows from Lemma E.2 and (ii) holds by (91). Then, we consider two cases w.r.t. the uncertainty level σ^+ to control (222), respectively:

- *When* $0 < \sigma^+ \le \frac{c_2}{2H}$*:* Recall $p = \frac{c_2}{H}$ if $\sigma^+ \le \frac{c_2}{2H}$. In this case, applying (222), we have

$$\begin{aligned}
& V_h^{\star,\sigma^+,f,\phi}(n_f) - V_h^{\star,\sigma^+,f,\phi}(m_f) \\
= & \sum_{j=0}^{H-h}(1-p)^j \ge \sum_{j=0}^{H-h}\left(1-\frac{c_2}{H}\right)^j = \frac{1-\left(1-\frac{c_2}{H}\right)^{H-h+1}}{c_2/H} \ge \frac{2c_2(H-h+1)}{3}. \qquad (223)
\end{aligned}$$

  Here, the final inequality holds by observing

$$\left(1-\frac{c_2}{H}\right)^{H-h+1} \le \exp\left(-\frac{c_2(H-h+1)}{H}\right) \le 1 - \frac{2c_2(H-h+1)}{3H}, \qquad (224)$$

  where the first inequality holds by noticing $c_2 < \frac{1}{2}$ and then $1-x \le \exp(-x)$, and the last inequality holds by $\exp(-x) \le 1 - \frac{2x}{3}$ for any $0 \le x \le \frac{1}{2}$.
  Plugging above fact in (223) back to (220), we arrive at

$$\begin{aligned}
& V_h^{\star,\sigma^+,f,\phi}(m_f) - V_h^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \\
\ge & (1-p^{\inf})\left(V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right) \\
& + \frac{1}{2}(p-q)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f) - \widetilde{\mu}_h(\cdot\,|\,m_f)\right\|_1 \frac{2c_2(H-h+1)}{3}. \qquad (225)
\end{aligned}$$

  Then, invoking the assumption

$$\sum_{h=1}^{H}\left\|\widetilde{\mu}_h(\cdot\,|\,m_f) - \widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f)\right\|_1 \ge \frac{H}{8} \qquad (226)$$

  in (111) and applying (225) recursively for $h = 1, 2, \cdots, H$ yields

$$\begin{aligned}
& V_1^{\star,\sigma^+,f,\phi}(m_f) - V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \\
\ge & \frac{c_2}{3}\sum_{h=1}^{H}(1-p^{\inf})^{h-1}(p-q)(H-h+1)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f) - \widetilde{\mu}_h(\cdot\,|\,m_f)\right\|_1 \\
\overset{\text{(i)}}{\ge} & \frac{c_2}{3}\sum_{h=1}^{H}(1-\frac{c_2}{H})^{h-1}(p-q)(H-h+1)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f) - \widetilde{\mu}_h(\cdot\,|\,m_f)\right\|_1 \\
\overset{\text{(ii)}}{\ge} & \frac{c_2}{6}\sum_{h=1}^{H}(p-q)(H-h+1)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot\,|\,m_f) - \widetilde{\mu}_h(\cdot\,|\,m_f)\right\|_1 \\
\overset{\text{(iii)}}{=} & \frac{c_2\Delta}{6}\sum_{h=1}^{H}h\left\|\widetilde{\mu}_{H-h+1}^{\star,f,\phi}(\cdot\,|\,m_f) - \widetilde{\mu}_{H-h+1}(\cdot\,|\,m_f)\right\|_1 \\
\overset{\text{(iv)}}{\ge} & \frac{c_2\Delta}{6}\sum_{h=1}^{\lfloor H/16\rfloor}2h \ge \frac{c_2\Delta}{6}\lfloor H/16\rfloor\left(\lfloor H/16\rfloor + 1\right), \qquad (227)
\end{aligned}$$

where (i) follows from $1-p^{\inf} \geq 1-p = 1-\frac{c_2}{H}$, and (ii) holds by

$$\forall h \in [H]: \quad (1-\frac{c_2}{H})^{h-1} \geq (1-\frac{c_2}{H})^{H} \geq \frac{1}{2}b \tag{228}$$

as long as $c_2 \leq \frac{1}{2}$. Here, (iii) arises from the definition of p, q in (99); (iv) can be verified by the fact that for any series $0 \leq m_1, m_2, \cdots, m_H \leq m_{\max}$ that obeys $\sum_{h=1}^{H} m_h \geq y$, one has

$$\sum_{h=1}^{H} m_h h \geq \sum_{h=1}^{\lfloor m_{\max}/n \rfloor} m_{\max} h, \tag{229}$$

and taking $m_h = \left\|\widetilde{\mu}_{H-h+1}(\cdot \,|\, m_f) - \widetilde{\mu}^{\star,f,\phi}_{H-h+1}(\cdot \,|\, m_f)\right\|_1 \leq 2 = m_{\max}$ and $n = \frac{H}{8}$. Consequently, observed from (227), the following inequality holds

$$V_1^{\star,\sigma^+,f,\phi}(m_f) - V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \geq \frac{c_2\Delta}{6}\lfloor H/16 \rfloor (\lfloor H/16 \rfloor + 1) \geq c_3 \Delta H^2 > \varepsilon \tag{230}$$

for some small enough constant c_3 and letting $\Delta = \frac{\varepsilon}{c_3 H^2}$.

- *When* $\frac{c_2}{2H} < \sigma^+ \leq 1 - c_0$*:* Similarly, recalling $p = \left(1+\frac{c_1}{H}\right)\sigma^+$ if $\sigma^+ > \frac{c_2}{2H}$ and invoking (222) gives

$$\begin{aligned} V_h^{\star,\sigma^+,f,\phi}(n_f) - V_h^{\star,\sigma^+,f,\phi}(m_f) &= \sum_{j=0}^{H-h}(1-p)^j = \sum_{j=0}^{H-h}\left(1-\left(1+\frac{c_1}{H}\right)\sigma^+\right)^j \\ &\geq \frac{1-\left(1-(1+\frac{c_1}{H})\sigma^+\right)^{H-h+1}}{(1+\frac{c_1}{H})\sigma^+} \\ &\geq \frac{c_2(H-h+1)}{3\sigma^+ H}, \end{aligned} \tag{231}$$

where the final inequality holds by observing

$$\begin{aligned} \left(1-\left(1+\frac{c_1}{H}\right)\sigma^+\right)^{H-h+1} &\leq \exp\left(-\left(1+\frac{c_1}{H}\right)\sigma^+(H-h+1)\right) \\ &\overset{\text{(i)}}{\leq} \exp\left(-\frac{c_2}{2H}\left(1+\frac{c_1}{H}\right)(H-h+1)\right) \\ &\leq 1-\left(1+\frac{c_1}{H}\right)\frac{c_2(H-h+1)}{3H}. \end{aligned} \tag{232}$$

Here, (i) holds by observing $\frac{c_2}{2H} < \sigma^+$, and the last inequality holds by $\left(1+\frac{c_1}{H}\right) \leq 2$, $c_2 \leq \frac{1}{2}$, and the fact $\exp(-x) \leq 1 - \frac{2x}{3}$ for any $0 \leq x \leq \frac{1}{2}$.
Plugging (231) into (220) gives

$$\begin{aligned} &V_h^{\star,\sigma^+,f,\phi}(m_f) - V_h^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \\ \geq &(1-p^{\inf})\left(V_{h+1}^{\star,\sigma^+,f,\phi}(m_f) - V_{h+1}^{\widetilde{\mu},\sigma^+,f,\phi}(m_f)\right) \\ &+ \frac{1}{2}(p-q)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot \,|\, m_f) - \widetilde{\mu}_h(\cdot \,|\, m_f)\right\|_1 \frac{c_2(H-h+1)}{3\sigma^+ H}. \end{aligned} \tag{233}$$

Following the steps to achieve (227), applying (233) recursively for $h = 1, 2, \cdots, H$ gives

$$\begin{aligned} &V_1^{\star,\sigma^+,f,\phi}(m_f) - V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \\ \geq &\sum_{h=1}^{H}(1-p^{\inf})^{h-1}(p-q)\frac{c_2(H-h+1)}{6\sigma^+ H}\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot \,|\, m_f) - \widetilde{\mu}_h(\cdot \,|\, m_f)\right\|_1 \\ \overset{\text{(i)}}{=} &\frac{c_2(p-q)}{6\sigma^+ H}\sum_{h=1}^{H}(1-\frac{c_1}{H})^{h-1}(H-h+1)\left\|\widetilde{\mu}_h^{\star,f,\phi}(\cdot \,|\, m_f) - \widetilde{\mu}_h(\cdot \,|\, m_f)\right\|_1 \\ \overset{\text{(ii)}}{\geq} &\frac{c_2\Delta}{12\sigma^+ H}\lfloor H/16 \rfloor (\lfloor H/16 \rfloor + 1), \end{aligned} \tag{234}$$

where (i) follows from $1 - p^{\inf} = 1 - (p - \sigma^+) = 1 - \frac{c_1}{H}\sigma^+$, and (ii) holds by letting $c_1 \leq \frac{1}{2}$ and following the same routine of (227). Consequently, (234) yields

$$V_1^{\star,\sigma^+,f,\phi}(m_f) - V_1^{\widetilde{\mu},\sigma^+,f,\phi}(m_f) \geq \frac{c_2\Delta}{12\sigma^+ H} \lfloor H/16 \rfloor (\lfloor H/16 \rfloor + 1) \geq \frac{c_4 \Delta H}{\sigma^+} > \varepsilon, \tag{235}$$

which holds for some small enough constant c_4 and letting $\Delta = \frac{\sigma^+ \varepsilon}{c_4 H}$.