Sample-Efficient Robust Multi-Agent Reinforcement Learning in the Face of Environmental Uncertainty

Laixi Shi [1] **Eric Mazumdar** [1] **Yuejie Chi** [2] **Adam Wierman** [1]

Abstract

To overcome the sim-to-real gap in reinforcement learning (RL), learned policies must maintain robustness against environmental uncertainties. While robust RL has been widely studied in single-agent regimes, in multi-agent environments, the problem remains understudied—despite the fact that the problems posed by environmental uncertainties are often exacerbated by strategic interactions. This work focuses on learning in distributionally robust Markov games (RMGs), a robust variant of standard Markov games, wherein each agent aims to learn a policy that maximizes its own worst-case performance when the deployed environment deviates within its own prescribed uncertainty set. This results in a set of robust equilibrium strategies for all agents that align with classic notions of game-theoretic equilibria. Assuming a non-adaptive sampling mechanism from a generative model, we propose a sample-efficient model-based algorithm (DR-NVI) with finite-sample complexity guarantees for learning robust variants of various notions of game-theoretic equilibria. We also establish an information-theoretic lower bound for solving RMGs, which confirms the near-optimal sample complexity of DR-NVI with respect to problem-dependent factors such as the size of the state space, the target accuracy, and the horizon length.

1. Introduction

Many real-world applications of artificial intelligence naturally involve multiple agents in dynamically evolving environments. Examples include ecosystem protection (Fang et al., 2015), board games (Silver et al., 2017), strategic management (Saloner, 1991), and autonomous driving (Zhou et al., 2020) among many others. One of the most promising algorithmic paradigms for addressing these problems is that of (deep) multi-agent reinforcement learning (MARL) (Silver et al., 2017; Vinyals et al., 2019; Lanctot et al., 2019) through a *decision-making* perspective. In full generality, it allows for agents with misaligned and possibly conflicting interests to optimize their own long-term rewards in an unknown dynamic environment, while taking one another into account. As such, MARL can often be modeled as learning in Markov games (MGs) (Littman, 1994; Shapley, 1953). Due to the game-theoretic nature of MGs, one often relies on solution concepts which take the form of equilibria — strategies/policies that are stable under rational deviations for all agents — like Nash equilibria (NE) (Nash, 1951; Shapley, 1953), correlated equilibria (CE) (Aumann, 1987), and coarse correlated equilibra (CCE) (Aumann, 1987; Moulin & Vial, 1978).

1.1. Environmental uncertainty in MARL

However, the equilibria of MGs can be very sensitive to environmental perturbations. Environmental uncertainties caused by system noise, model mismatch, and sim-to-real gaps can cause dramatic changes to both the qualitative outcomes of the game as well as agents' payoffs. While this problem is present in single-agent RL, the need for robustness is even more acute in the multi-agent setting where the game-theoretic interactions can cause instabilities (Slumbers et al., 2023). Indeed, playing an equilibrium solution learned in the simulated environment might lead to a catastrophic drop in a single agent's payoff or even all agents' payoffs when the deployed environment deviates slightly from what is expected (Balaji et al., 2019; Zhang et al., 2020c; Zeng et al., 2022; Yeh et al., 2021), a point we illustrate in the following example.

Example: fishing protection. *To emphasize the impact of model uncertainty in MARL, in Figure 1, we present a concrete example of a simple two-player game that models the interaction between a fisherman and law enforcement trying to prevent illegal fishing. The state* $s \in \{0, 1, \cdots, 100\}$ *rep-*

[1]Department of Computing Mathematical Sciences, California Institute of Technology, CA 91125, USA. [2]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.. Correspondence to: Laixi Shi <laixis@caltech.edu>.

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

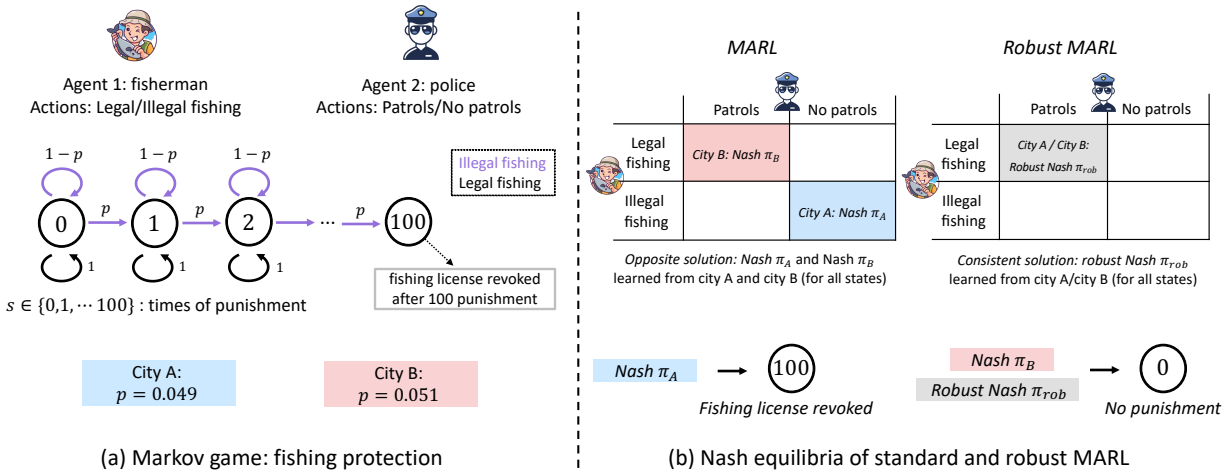


Figure 1. A two-player general-sum Markov game modeling preventing illegal fishing. (a) shows the state space (circles) and the simplified transitions; the fisherman arrives at distinct states by executing different Nash equilibrium solutions π_A (from city A) or π_B (from city B). (b) in two slightly different environments (city A versus city B), it shows the solutions π_A, π_B of the standard game, and the consistent solution *robust Nash* π_{rob} of a robust variant of the game (detailed in Appendix B.1).

resents the number of punishments received by the fisherman, with the license being revoked at $s = 100$. The environment is governed by a model parameter p. We observe from Figure 1(b) that for slightly perturbed environments, city A ($p = 0.049$) and city B ($p = 0.051$), the solutions of the MGs are two Nash equilibria with drastically different outcomes: no punishment under policy π_B learned from city B (in red) and a revoked license under policy π_A learned from city A (in blue). More details are presented in Appendix B.1. The example above illustrates how the standard formulation of a MG can be vulnerable to model uncertainties and result in unstable solutions with divergent outcomes. As such, robustness and stability become a pressing need and key challenge for the deployment of MARL algorithms.

To address this, we consider robust MARL problems as (distributionally) robust Markov games (RMGs) — a robust counterpart of standard MGs (Zhang et al., 2020c; Kardeş et al., 2011). The natural solution concepts for RMGs are equilibria not only between agents, but also between multiple natural adversaries that choose the worst-case environments within some prescribed uncertainty set for each agent. By design, they exhibit more robustness and consistency in the face of unmodeled disturbances. To illustrate this, consider the example in Figure 1, where one can observe that the solutions of a RMG (π_{rob} in gray) remains consistent and stable across similar environments city A and city B.

Despite some recent efforts (Zhang et al., 2020c; Kardeş et al., 2011; Ma et al., 2023; Blanchet et al., 2023), a fundamental understanding of learning in RMGs is lacking. Indeed, while the robust formulation of single-agent RL has been well studied (Iyengar, 2005; Nilim & El Ghaoui, 2005; Shi et al., 2023; Xu et al., 2023), understanding how to efficiently learn equilibrium policies in robust Markov games remains an open question. We focus on understanding and achieving near-optimal sample efficiency in robust MGs, reflecting the fact that in many large-scale applications, agents must learn from samples from an unknown but potentially extremely large environment (Silver et al., 2016; Vinyals et al., 2019; Achiam et al., 2023). While some attempts have been made to design sample-efficient algorithms for robust MARL (Wang et al., 2023a; Blanchet et al., 2023), the current solutions are still far from optimal. With that in mind, we investigate the following open question:

> *Can we achieve robustness and near-optimal sample efficiency in MARL simultaneously?*

1.2. Main contributions

To address the open question, this work concentrates on designing algorithms for robust MGs with near-optimal sample complexity guarantees. We consider three solution concepts for RMGs, which are robust variants of standard equilibria — robust NE, robust CE, and robust CCE. We focus on a class of RMGs, where the uncertainty sets of the environment are constructed following an *agent-wise (s,a)-rectangularity* condition for computational tractability (Iyengar, 2005; Wiesemann et al., 2013) (see Section 3). Such a condition allows each agent to independently consider its uncertainty set according to their personal interest. We consider total variation (TV) distance as the distance metric for the uncertainty set, motivated by its practical (Pan et al., 2023; Lee et al., 2021) and theoretical appeal (Panaganti & Kalathil, 2022; Shi et al., 2023; Blanchet et al., 2023).

Concretely, our study focuses on finite-horizon RMGs with n agents. We denote the episode length by H, the size of the state space by S, the size of the i-th agent's action space by A_i, and use $\sigma_i \in (0,1]$ to represent the uncertainty level of the i-th agent. We assume access to a generative model that can draw samples from the nominal environment in a non-adaptive manner. The goal is to find an ε-approximate equilibrium for RMGs — a joint policy such that each agent's benefit is at most ε away under rational deviations. The main contributions are summarized as follows.

- *Near-optimal sample complexity upper bound.* We design a model-based algorithm — distributionally robust Nash value iteration (DR-NVI), which can provably find any solution among ε-approximate robust-{NE, CCE, CE} with high probability, when the sample size exceeds

$$\widetilde{O}\left(\frac{SH^3\prod_{i=1}^n A_i}{\varepsilon^2}\min\left\{H,\ \frac{1}{\min_{1\le i\le n}\sigma_i}\right\}\right). \quad (1)$$

  This significantly improves upon prior art (Blanchet et al., 2023) $\widetilde{O}\big(S^4\left(\prod_{i=1}^n A_i\right)^3 H^4/\varepsilon^2\big)$[1] (Blanchet et al., 2023) by at least a factor of $\widetilde{O}\big(S^3\left(\prod_{i=1}^n A_i\right)^2\big)$, and further delineates the impact of the uncertainty levels. Our results are derived by addressing the intricate statistical dependencies arising from game-theoretical interactions among agents, a challenge not present in robust single-agent RL. Additionally, we employ distributionally robust optimization to address the nonlinear payoffs of agents in RMGs, which lack a closed form.

- *Information-theoretic lower bound.* To understand the optimality of our algorithm we establish a lower bound for solving RMGs, showing that no algorithm can learn any of ε-approximate robust-{NE, CCE, CE} with fewer samples than

$$\widetilde{O}\left(\frac{SH^3\max_{1\le i\le n}A_i}{\varepsilon^2}\min\left\{H,\ \frac{1}{\min_{1\le i\le n}\sigma_i}\right\}\right). \quad (2)$$

  To the best of our knowledge, this is the first information-theoretic lower bound for RMGs, regardless of the distance metric in use. We construct new hard scenarios for tightness, differing from existing ones in both robust single-agent RL and standard MGs, which may be of independent interest. This in turn establishes that the sample complexity of DR-NVI is optimal for all RMGs with respect to many critical problem-dependent parameters such as $S, H, \{\sigma_i\}_{1\le i\le n}$, making DR-NVI the first near-optimal finite-sample guarantee for robust MGs, regardless of the divergence metric in use.

[1]Note that Blanchet et al. (2023) targets a different (and more challenging) setting with offline data. We translate the results of Blanchet et al. (2023) to the generative setting we consider.

Notation. Throughout this paper, we introduce the notation $[T] \coloneqq \{1,\cdots,T\}$ for any positive integer $T>0$. We denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$ and $x = \big[x(s,a)\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{SA}$ (resp. $x = \big[x(s)\big]_{s\in\mathcal{S}} \in \mathbb{R}^S$) as any vector that constitutes certain values for each state-action pair (resp. state).

2. Background: Standard Markov Games

We begin by covering the foundational aspects of multi-agent general-sum standard Markov games in a finite-horizon setting.

Standard Markov games. A finite-horizon *multi-agent general-sum Markov game* can be represented as $\mathcal{MG} = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, P, r, H\big\}$. This game involves n agents who optimizes their own benefits in a shared environment, consisting of the following key components.

- State space $\mathcal{S} = \{1,\cdots,S\}$ of the shared environment with S different states.
- Joint action space $\mathcal{A}$: for each $1\le i\le n$, we represent $\mathcal{A}_i = \{1,\cdots,A_i\}$ as the action space of the i-th agent that contains A_i different actions. In addition, we denote the joint action space for all agents (or a subset of agents) as $\mathcal{A} \coloneqq \mathcal{A}_1\times\cdots\times\mathcal{A}_m$ (or $\mathcal{A}_{-i} \coloneqq \prod_{j:j\neq i}\mathcal{A}_j$ for all $1\le i\le n$). For convenience, we denote the boldface letter $\boldsymbol{a}\in\mathcal{A}$ (resp. $\boldsymbol{a}_{-i}\in\mathcal{A}_{-i}$) as a joint action profile for all agents (resp. all agents excluding the i-th agent).
- Probability transition kernel $P = \{P_h\}_{1\le h\le H}$ with $P_h : p$). Specifically, $P_h(s' \mid s, \boldsymbol{a})$ represents the probability of $\mathcal{MG}$ transitioning from current state $s\in\mathcal{S}$ to the next state $s'\in\mathcal{S}$ at time step h, given the agents choose the joint action profile $\boldsymbol{a}\in\mathcal{A}$.
- Reward function $r = \{r_{i,h}\}_{1\le i\le n, 1\le h\le H}$ with $r_{i,h} : \mathcal{S}\times\mathcal{A}\mapsto[0,1]$. Specifically, for any $(i,h,s,\boldsymbol{a}) \in [n]\times[H]\times\mathcal{S}\times\mathcal{A}$, let $r_{i,h}(s,\boldsymbol{a})$ be the immediate (deterministic) reward received by the i-th agent in state s when the joint action profile is $\boldsymbol{a}$, which is normalized to $[0,1]$ without loss of generality.
- H is the horizon length of the standard MG.

Markov policies and value functions. Throughout the paper, we focus on the class of Markov policies, namely, the action selection rule is solely determined by the current state s, independent from previous trajectories (including

visited states, executed actions, and received rewards) of all agents. Specifically, for any $1 \le i \le n$, the i-th agent executes actions according to a policy $\pi_i = \{\pi_{i,h} : \mathcal{S} \mapsto \Delta(\mathcal{A}_i)\}_{1\le h\le H}$, with $\pi_{i,h}(a \,|\, s)$ the probability of selecting action a in state s at time step h. The joint Markov policy of all agents can be defined as $\pi = (\pi_1, \ldots, \pi_n) : \mathcal{S} \times [H] \mapsto \Delta(\mathcal{A})$, namely, the joint action profile $\boldsymbol{a}$ of all agents is chosen according to the distribution specified by $\pi_h(\cdot \,|\, s) = (\pi_{1,h}, \pi_{2,h} \ldots, \pi_{n,h})(\cdot \,|\, s) \in \Delta(\mathcal{A})$ conditioned on state s at time step h.

With the above notation in mind, for any given joint policy π and transition kernel P of the $\mathcal{MG}$, we characterize the long-term cumulative reward by defining the value function $V_{i,h}^{\pi,P} : \mathcal{S} \mapsto \mathbb{R}$ (resp. Q-function $Q_{i,h}^{\pi,P} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$) of the i-th agent as follows: for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$V_{i,h}^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_{i,t}\big(s_t, \boldsymbol{a}_t\big) \mid s_h = s\right],$$
$$Q_{i,h}^{\pi,P}(s, \boldsymbol{a}) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_{i,t}\big(s_t, \boldsymbol{a}_t\big) \mid s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\right], \tag{3}$$

where the expectation is taken over the Markovian trajectory $\{(s_t, \boldsymbol{a}_t)\}_{h\le t\le H}$ by executing the joint policy π under the transition kernel P, i.e., $\boldsymbol{a}_t \sim \pi_t(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, \boldsymbol{a}_t)$.

Best-response policy. For any given joint policy π, we employ π_{-i} to represent the policies of all agents excluding the i-th agent. We define the maximum value function of the i-th agent at time step h against the joint policy π_{-i} of the other agents as

$$V_{i,h}^{\star,\pi_{-i},P}(s) \coloneqq \max_{\pi_i' : \mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} V_{i,h}^{\pi_i'\times\pi_{-i},P}(s), \tag{4}$$

where $\pi_i' \times \pi_{-i}$ represents the joint policy of all agents when the i-th agent executes policy π_i'. It is well-known (Filar & Vrieze, 2012) that there exists at least one Markovian policy, the *best-response policy*, that achieves $V_{i,h}^{\star,\pi_{-i},P}(s)$ for all $s \in \mathcal{S}$ and all $h \in [H]$ simultaneously. We denote the best-response policy using $\pi_i^{\star,P}(\pi_{-i}) : \mathcal{S} \times [H] \mapsto \Delta(\mathcal{A}_i)$.

Solution concepts: equilibria. In MGs, strategic agents are modeled in a possibly competitive framework and focus on finding some sort of equilibrium strategies. Here, we consider three common types of equilibria — NE, CE, and CCE for MGs.

- *Nash equilibrium (NE).* A product policy $\pi = \pi_1 \times \cdots \times \pi_n \in \Delta(\mathcal{A}_1) \times \Delta(\mathcal{A}_2) \times \cdots \times \Delta(\mathcal{A}_n)$ is said to be a *(mixed-strategy Markov) NE* if

$$\text{for all } (s, i) \in \mathcal{S} \times [n] : \; V_{i,1}^{\pi,P}(s) = V_{i,1}^{\star,\pi_{-i},P}(s). \tag{5}$$

  Namely, as long as all players act independently, no player can benefit by unilaterally diverging from its present policy, given the current policies of the opponents.

- *Coarse correlated equilibrium (CCE).* A joint policy $\pi \in \Delta(\mathcal{A})$ is said to be a CCE (Moulin & Vial, 1978; Aumann, 1987) if it holds that

$$\text{for all } (s, i) \in \mathcal{S} \times [n] : \; V_{i,1}^{\pi,P}(s) \ge V_{i,1}^{\star,\pi_{-i},P}(s). \tag{6}$$

  As a relaxation of NE, CCE also guarantees that no player has incentive to unilaterally deviated from the current policy. The key difference from the NE definition is that it permits policies to be interrelated among players.

- *Correlated equilibrium (CE).* Before proceeding, for each $1 \le i \le n$, we define a set of function $f_i \coloneqq \{f_{i,h,s}\}_{h\in[H],s\in\mathcal{S}}$ with $f_{i,h,s} : \mathcal{A}_i \mapsto \mathcal{A}_i$, and denoting $\mathcal{F}_i$ as the set of possible f_i. Armed with this, we can combine such f_i with any joint policy π to reach a new policy $f_i \diamond \pi$, where $f_i \diamond \pi$ will choose $(a_1, \ldots, a_{i-1}, f_i(a_i), a_{i+1}, \ldots, a_n)$ when policy π selects $(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n)$. With these in place, a joint policy $\pi \in \Delta(\mathcal{A})$ is said to be a CE (Moulin & Vial, 1978; Aumann, 1987) if it holds that

$$\text{for all } (s, i) \in \mathcal{S}\times[n] : \; V_{i,1}^{\pi,P}(s) \ge \max_{f_i\in\mathcal{F}_i} V_{i,1}^{f_i\diamond\pi,P}(s). \tag{7}$$

  CE is a also a relaxation of NE, which does not require the joint policy π to be a product policy.

3. Distributionally Robust Markov Games

We consider a robust variant of standard MGs incorporating environmental uncertainties — termed *distributionally robust Markov games* (RMGs). RMGs represent a richer class than standard MGs, allowing for different prescribed environmental uncertainty sets as long as they meet a *rectangularity* condition, detailed below.

3.1. Distributionally robust Markov games

A *distributionally robust multi-agent general-sum Markov game* (RMG) in the finite-horizon setting is defined by

$$\mathcal{MG}_{\mathsf{rob}} = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, \{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{1\le i\le n}, r, H\big\},$$

where $\mathcal{S}, \{\mathcal{A}_i\}, r$, and H are identical to those of standard MGs (see Section 2). A notable deviation from standard MGs is that: for $1 \le i \le n$, instead of assuming a fixed transition kernel, each i-th agent 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 a *nominal* kernel

$P^0 : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$, with its size and shape defined by a certain distance metric ρ and a radius parameter $\sigma_i > 0$. Note that, for generality, to accommodate individual robustness preferences, each agent is permitted to tailor its own uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$ by choosing different size σ_i and even the shape determined by different divergence function ρ. Here, we consider the same divergence function for all agents for simplicity. And we focus on the discussion of the transition kernel's uncertainty in this work, it's worth noting that similar uncertainty can also be considered for each agent's reward function.

Uncertainty set with *agent-wise* (s,a)*-rectangularity*. In the following, we specify the construction of the transition kernel uncertainty sets $\mathcal{U}_\rho(P^0) = \{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{1\le i\le n}$ for RMGs. Drawing inspiration from the *rectangularity* condition advocated in robust single-agent RL (Iyengar, 2005; Wiesemann et al., 2013; Zhou et al., 2021; Shi et al., 2023), we consider a multi-agent variant of rectangularity in RMGs — *agent-wise* (s,a)*-rectangularity*. This condition enables the robust counterpart of Bellman recursions and computational tractability of the problems. It allows for each agent to independently choose its own uncertainty set that can be decomposed into a product of subsets over each state-action pair.

In particular, we assume all agents use the same distance metric ρ for their uncertainty sets.[2] Each i-th agent can choose their own uncertainty level $\sigma_i > 0$ independently. With ρ and $\{\sigma_i\}_{1\le i\le n}$ in hand, the uncertainty set $\mathcal{U}_\rho(P^0)$ of all agents obeying *agent-wise* (s,a)*-rectangularity* is mathematically specified as: for all $i \in [n]$,

$$\mathcal{U}_\rho^{\sigma_i}(P^0) := \otimes\, \mathcal{U}^{\sigma_i}(P^0_{h,s,\boldsymbol{a}}) \quad \text{with} \tag{8}$$
$$\mathcal{U}_\rho^{\sigma_i}(P^0_{h,s,\boldsymbol{a}}) := \left\{P_{h,s,\boldsymbol{a}} \in \Delta(\mathcal{S}) : \rho\left(P_{h,s,\boldsymbol{a}}, P^0_{h,s,\boldsymbol{a}}\right) \le \sigma_i\right\},$$

where $\otimes$ represents the Cartesian product and we denote a vector of the transition kernel P or P^0 at any state-action pair $(s,\boldsymbol{a}) \in \mathcal{S}\times\mathcal{A}$ respectively as

$$P_{h,s,\boldsymbol{a}} := P_h(\cdot \mid s,\boldsymbol{a}) \in \mathbb{R}^{1\times S},$$
$$P^0_{h,s,\boldsymbol{a}} := P^0_h(\cdot \mid s,\boldsymbol{a}) \in \mathbb{R}^{1\times S}. \tag{9}$$

Here, the 'distance' function ρ for each agent's uncertainty set can be chosen from many candidate functions that measure the difference between two probability vectors, such as f-divergence (including total variation (TV), chi-square, and Kullback-Leibler (KL) divergence) (Yang et al., 2022), ℓ_q norm (Clavier et al., 2023), and Wasserstein distance (Xu et al., 2023). In this work, we focus on the uncertainty sets that are constructed using TV distance:

$$\rho_{\mathsf{TV}}\left(P_{h,s,\boldsymbol{a}}, P^0_{h,s,\boldsymbol{a}}\right) := \frac{1}{2}\left\|P_{h,s,\boldsymbol{a}} - P^0_{h,s,\boldsymbol{a}}\right\|_1. \tag{10}$$

[2] Generally, each agent can decide their own (possibly different) distance metric for the uncertainty set. We consider the same ρ for simplicity.

Robust value functions. For a RMG, each agent aims to maximize its own worst-case performance over all possible transition kernels in its own (possibly different) prescribed uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$. For any joint policy $\pi \in \Delta(\mathcal{A})$, the worst-case performance of the i-th agent at time step h can be measured by the *robust value function* $V_{i,h}^{\pi,\sigma_i}$ and the *robust Q-function* $Q_{i,h}^{\pi,\sigma_i}$, defined as

$$V_{i,h}^{\pi,\sigma_i}(s) := \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P^0)} V_{i,h}^{\pi,P}(s)$$
$$Q_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a}) := \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P^0)} Q_{i,h}^{\pi,P}(s,\boldsymbol{a}) \tag{11}$$

for all $(i,h,s,\boldsymbol{a}) \in [n]\times[H]\times\mathcal{S}\times\mathcal{A}$. Similar to standard MGs, given a fixed joint policy π_{-i} for all agents but the i-th agent, by optimizing over $\pi_i' : \mathcal{S}\times[H] \to \Delta(\mathcal{A}_i)$ that is executed independently from π_{-i}, we can further define the maximum of the robust value function for each agent as follows: for all $(i,h,s) \in [n]\times[H]\times\mathcal{S}$:

$$V_{i,h}^{\star,\pi_{-i},\sigma_i}(s) := \max_{\pi_i':\mathcal{S}\times[H]\mapsto\Delta(\mathcal{A}_i)} V_{i,h}^{\pi_i'\times\pi_{-i},\sigma_i}(s)$$
$$= \max_{\pi_i':\mathcal{S}\times[H]\mapsto\Delta(\mathcal{A}_i)} \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P^0)} V_{i,h}^{\pi_i'\times\pi_{-i},P}(s). \tag{12}$$

Similar to standard MGs, it can be easily verified that there exists at least one policy (Blanchet et al., 2024, Section A.2), denoted by $\pi_i^{\star,\sigma_i}(\pi_{-i}) : \mathcal{S}\times[H] \to \Delta(\mathcal{A}_i)$ and referred to as the *robust best-response policy* for the i-th agent, that can simultaneously attain $V_{i,h}^{\star,\pi_{-i},\sigma_i}(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$.

Robust Bellman equations. Analogous to standard MGs, RMGs feature a robust counterpart of the Bellman equation — *robust Bellman equation*. In particular, the robust value functions $\{V_{i,h}^{\pi,\sigma_i}\}$ of RMGs associated with any joint policy π obey: for all $(i,h,s) \in [n]\times[H]\times\mathcal{S}$,

$$V_{i,h}^{\pi,\sigma_i}(s)$$
$$= \mathbb{E}_{\boldsymbol{a}\sim\pi_h(s)}\left[r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})} PV_{i,h+1}^{\pi,\sigma_i}\right]. \tag{13}$$

We emphasize that the above robust Bellman equation is fundamentally linked to the *agent-wise* (s,a)*-rectangularity* condition (cf. (8)) imposed on the designed uncertainty set. Specifically, this condition decouples the dependency of uncertainty subsets across different agents, each state-action pair, and different time steps, leading to the Bellman recursive equation.

3.2. Solution concepts for robust Markov games

For RMGs, the games are no longer n-agent games, but become $2n$-agent games between agents and n natural adversaries to choose the worst-case transitions. Given the

possibly conflicting objectives, finding an equilibrium becomes a core goal for RMGs. Below, we introduce three robust variants of widely considered standard solution concepts — robust NE, robust CE, and robust CCE for any RMG.

- *Robust NE.* A product policy $\pi = \pi_1 \times \pi_2 \times \cdots \times \pi_n$ is said to be a *robust NE* if (cf. (5))

$$\forall (i,s) \in [n] \times \mathcal{S}: \quad V_{i,1}^{\pi,\sigma_i}(s) = V_{i,1}^{\star,\pi_{-i},\sigma_i}(s). \tag{14}$$

  Robust NE indicates that given the current strategy of the opponents π_{-i}, when each agent considers the worst-case performance over its own uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$, no player can benefit by unilaterally diverging from its present strategy.

- *Robust CCE.* A (possibly correlated) joint policy $\pi \in \mathcal{S} \times [H] \mapsto \Delta(\mathcal{A})$ is said to be a *robust CCE* if it holds that (cf. (6))

$$\forall (i,s) \in [n] \times \mathcal{S}: \; V_{i,1}^{\pi,\sigma_i}(s) \geq V_{i,1}^{\star,\pi_{-i},\sigma_i}(s). \tag{15}$$

  As a relaxation of robust NE, robust CCE also guarantees that no player has incentive to unilaterally deviate from the current policy, where the policies are not necessarily independent among players.

- *Robust CE.* A joint policy $\pi \in \Delta(\mathcal{A})$ is said to be a robust CE if it holds that (cf. (7))

$$\forall (s,i) \in \mathcal{S} \times [n]: \; V_{i,1}^{\pi,\sigma_i}(s) \geq \max_{f_i \in \mathcal{F}_i} V_{i,1}^{f_i \diamond \pi,\sigma_i}(s). \tag{16}$$

It is known that computing exact robust equilibria is challenging and may not be necessary in practice. As a result, people usually search for approximate equilibria. Toward this, as a slightly relaxation from (14), a product policy $\pi \in \Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_n)$ is said to be an *ε-robust NE* if

$$\mathsf{gap}_{\mathsf{NE}}(\pi) := \max_{s \in \mathcal{S}, 1 \leq i \leq n} \left\{ V_{i,1}^{\star,\pi_{-i},\sigma_i}(s) - V_{i,1}^{\pi,\sigma_i}(s) \right\} \leq \varepsilon. \tag{17}$$

Similarly, relaxing (15) or (16), a (possibly correlated) joint policy $\pi \in \Delta(\mathcal{A})$ is said to be an *ε-robust CCE* if

$$\mathsf{gap}_{\mathsf{CCE}}(\pi) := \max_{s \in \mathcal{S}, 1 \leq i \leq n} \left\{ V_{i,1}^{\star,\pi_{-i},\sigma_i}(s) - V_{i,1}^{\pi,\sigma_i}(s) \right\} \leq \varepsilon, \tag{18}$$

or an *ε-robust CE* if

$$\mathsf{gap}_{\mathsf{CE}}(\pi) := \max_{s \in \mathcal{S}, 1 \leq i \leq n} \left\{ \max_{f_i \in \mathcal{F}_i} V_{i,1}^{f_i \diamond \pi,\sigma_i}(s) - V_{i,1}^{\pi,\sigma_i}(s) \right\} \leq \varepsilon. \tag{19}$$

The existence of robust NE has been verified (Blanchet et al., 2023) under general divergence functions for the uncertainty set. Indeed, the robust equilibria defined here can be reduced to the standard equilibria associated with the robust variant of standard payoffs (robust Q-functions), which have been verified obeying $\{\text{NE}\} \subseteq \{\text{CE}\} \subseteq \{\text{CCE}\}$ (Roughgarden, 2010). Therefore, the existence of robust NE directly indicates the existence of robust CE and robust CCE.

3.3. Non-adaptive sampling from a generative model

Given the formulation of distributionally robust Markov games, a question of prime interest is how to learn the robust equilibria without knowing the model exactly in a sample-efficient manner.

Sampling mechanism: a generative model. As a widely used sampling mechanism in standard MARL (Zhang et al., 2020b; Li et al., 2022a), in this paper, we assume access to a generative model (simulator) (Kearns & Singh, 1999) and collect samples in a non-adaptive manner. Specifically, for each tuple $(s, \boldsymbol{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we collect N independent samples generated based on the true *nominal* transition kernel P^0:

$$s_{i,h,s,\boldsymbol{a}} \overset{i.i.d}{\sim} P_h^0(\cdot \,|\, s, \boldsymbol{a}), \qquad i = 1, 2, \ldots, N. \tag{20}$$

The total number of samples is thus $N_{\mathsf{all}} = NS \prod_{i=1}^n A_i$.

Armed with the collected dataset from the nominal environment, the goal is to learn a solution among ε-robust-{NE, CCE, CE} for the game $\mathcal{MG}_{\mathsf{rob}}$ — w.r.t. some prescribed uncertainty set $\mathcal{U}(P^0)$ around the nominal kernel — using as few samples as possible.

4. Algorithm and Theory

In this and the following sections, we focus on the class of robust MGs with uncertainty set measured by TV distance, namely, the uncertainty set $\mathcal{U}_\rho^{\sigma_i}(\cdot) = \mathcal{U}_{\rho\mathsf{TV}}^{\sigma_i}(\cdot)$ w.r.t the TV distance $\rho = \rho_{\mathsf{TV}}$ defined in (10). For convenience, we abbreviate $\mathcal{U}^{\sigma_i}(\cdot) := \mathcal{U}_{\rho\mathsf{TV}}^{\sigma_i}(\cdot)$.

4.1. Distributionally robust Nash value iteration

We develop a model-based approach tailored to solve robust Markov games, which involves two separate steps. First, we construct an empirical nominal transition kernel $\widehat{P}^0$ using the collected samples from the generative model. Then armed with $\widehat{P}^0$, we propose to apply distributionally robust Nash value iteration (DR-NVI) to compute a robust equilibrium solution for all agents.

Nominal model estimation. Based on the empirical frequency of state transitions, we estimate the empirical nom-

inal transition kernel $\widehat{P}^0 = \{\widehat{P}^0_h\}_{h\in[H]}$, where the entries of $\widehat{P}^0_h \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$ at each time step h is constructed as follows: for all $(h, s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$,

$$\widehat{P}^0_h(s' \,|\, s, \boldsymbol{a}) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{s_{i,h,s,\boldsymbol{a}} = s'\}. \tag{21}$$

Distributionally robust Nash value iteration (DR-NVI). With the empirical nominal kernel $\widehat{P}^0$ in hand, to compute a robust equilibrium solution, we propose DR-NVI by adapting a model-based algorithm for standard Markov games — Nash value iteration (Liu et al., 2021), summarized in Algorithm 1.

The process starts from the last time step $h = H$ and proceeds with $h = H-1, H-2, \cdots, 1$. At each time step $h \in [H]$, the robust Q-function can be estimated as $\widehat{Q}_{i,h}$ (see line 4.1) as: for all $(i, h, s, \boldsymbol{a}) \in [n] \times [H] \times \mathcal{S} \times \mathcal{A}$,

$$\widehat{Q}_{i,h}(s, \boldsymbol{a}) = r_{i,h}(s, \boldsymbol{a}) + \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} \mathcal{P}\widehat{V}_{i,h+1}. \tag{22}$$

Directly solving (22) presents significant computational challenges due to the need to optimize over an S-dimensional probability simplex, a task whose complexity increases exponentially with the state space size S. Fortunately, leveraging strong duality enables us to solve (22) equivalently via its dual problem (Iyengar, 2005):

$$\widehat{Q}_{i,h}(s, \boldsymbol{a}) = r_{i,h}(s, \boldsymbol{a}) + \max_{\alpha\in[\min_s \widehat{V}_{i,h+1}(s), \max_s \widehat{V}_{i,h+1}(s)]} \left\{\widehat{P}^0_{h,s,\boldsymbol{a}}\left[\widehat{V}_{i,h+1}\right]_\alpha - \sigma_i\left(\alpha - \min_{s'}\left[\widehat{V}_{i,h+1}\right]_\alpha(s')\right)\right\}, \tag{23}$$

where $[V]_\alpha$ denotes the clipped version of any vector $V \in \mathbb{R}^S$ determined by some level $\alpha \geq 0$, namely,

$$[V]_\alpha(s) := \begin{cases} \alpha, & \text{if } V(s) > \alpha, \\ V(s), & \text{otherwise.} \end{cases} \tag{24}$$

With robust Q-function estimates $\{\widehat{Q}_{i,h}\}_{i\in[n]}$ available for all agents at time step h, the sub-routine in line 4.1 $\mathsf{Equilibrium} \in \mathsf{Compute}-\{\mathsf{Nash}, \mathsf{CE}, \mathsf{CCE}\}$ represents the algorithm for computing the corresponding robust-{NE, CE, CCE}, respectively. Note that for the studied RMGs, a robust-NE/CE/CCE is equivalent to a corresponding NE/CE/CCE associated with the payoff matrices $\{\widehat{Q}_{i,h}\}_{i\in[n]}$. On the computing and learning front of the sub-routine $\mathsf{Equilibrium}(\cdot)$, for a general standard MG, the NE has been proved PPAD-hard to compute (Daskalakis, 2013), even for two-player matrix games (except for two-player zero-sum games). Notably, even when the non-robust standard MG associated with the nominal transition kernel is a two-player zero-sum game, the corresponding robust MG is generally not because agents may select different worst-case transition kernels. Conversely, computing CE/CCE is computationally tractable within polynomial time through linear programming (Liu et al., 2021).

Algorithm 1 Distributionally robust equilibrium value iteration (DR-NVI).

input: empirical nominal transition kernel $\widehat{P}^0$; reward function r; uncertainty levels $\{\sigma_i\}_{i\in[n]}$.
initialization: $\widehat{Q}_{i,h}(s,a) = 0$, $\widehat{V}_{i,h}(s) = 0$ for all $(s, \boldsymbol{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H+1]$.
for $h = H, H-1, \cdots, 1$ **do**
  for $i = 1, 2, \cdots, n$ and $s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}$ **do**
    Set $\widehat{Q}_{i,h}(s, \boldsymbol{a})$ according to (22).
  end for
  for $s \in \mathcal{S}$ **do**
    Get $\pi_h(s) = \{\pi_{i,h}(s)\}_{1\leq i\leq n} \leftarrow \mathsf{Equilibrium}\left(\{\widehat{Q}_{i,h}(s,\cdot)\}_{1\leq i\leq n}\right)$.
    Set $\widehat{V}_{i,h}(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi_h}[\widehat{Q}_{i,h}(s, \boldsymbol{a})]$.
  end for
end for
output: $\{\widehat{Q}_{i,h}\}$, $\{\widehat{V}_{i,h}\}$, and $\widehat{\pi} = \{\pi_h\}_{1\leq h\leq H}$.

4.2. Sample complexity: upper and lower bounds

We now present our main theoretical results regarding the sample complexity of learning robust equilibria of robust Markov games, including an upper bound of DR-NVI (Algorithm 1) and an information-theoretic lower bound. First, we introduce the finite-sample guarantee for DR-NVI, which is proven in Appendix C.

Theorem 4.1 (Upper bound for DR-NVI). *Recall the TV uncertainty set $\mathcal{U}^{\sigma_i}(\cdot) = \mathcal{U}^{\sigma_i}_{\rho\mathsf{TV}}(\cdot)$ defined in (9). Consider any $\delta \in (0,1)$ and any RMG $\mathcal{MG}_{\mathsf{rob}} = \{\mathcal{S}, \{\mathcal{A}_i\}_{1\leq i\leq n}, \{\mathcal{U}^{\sigma_i}(P^0)\}_{1\leq i\leq n}, r, H\}$ with $\sigma_i \in (0,1]$ for all $i \in [n]$. For any $\varepsilon \leq \sqrt{\min\{H, \frac{1}{\min_{1\leq i\leq n}\sigma_i}\}}$, Algorithm 1 can output any robust equilibrium among ε-robust {NE, CCE, CE} by executing different sub-routine $\mathsf{Equilibrium} \in \mathsf{Compute}-\{\mathsf{Nash}, \mathsf{CE}, \mathsf{CCE}\}$ in line 4.1. Namely, for some constant C_1 and $\xi := \log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)$, we can achieve any of the following results*

$$\mathsf{gap}_{\mathsf{NE}}(\widehat{\pi}) \leq \varepsilon,$$
$$\mathsf{gap}_{\mathsf{CCE}}(\widehat{\pi}) \leq \varepsilon,$$
$$\mathsf{gap}_{\mathsf{CE}}(\widehat{\pi}) \leq \varepsilon$$

with probability at least $1-\delta$, as long as the total number of samples obeys

$$N_{\mathsf{all}} \geq \frac{C_1\xi SH^3\prod_{1\leq i\leq n} A_i}{\varepsilon^2}\min\left\{H, \frac{1}{\min_{1\leq i\leq n}\sigma_i}\right\}.$$

Before delving into the implications of Theorem 4.1, we provide a lower bound for solving robust Markov games. The proof is provided in Appendix D.

Theorem 4.2 (Lower bound for solving robust MGs)**.** *Consider any tuple* $\big\{S, \{A_i\}_{1\le i\le n}, \{\sigma_i\}_{1\le i\le n}, H\big\}$ *obeying* $\sigma_i \in (0, 1-c_0]$ *with* $0 < c_0 \le \frac{1}{4}$ *being any small enough positive constant, and* $H > 16\log 2$*. Let*

$$\varepsilon \le \begin{cases} \frac{c_0}{2H}, & \text{if } \sigma_1 \le \frac{c_0}{4H}, \\ 1 & \text{otherwise} \end{cases} \tag{25}$$

We can construct a set of RMGs— denoted as $\mathcal{M} = \{\mathcal{MG}_i\}_{i\in[I]}$*, such that for any dataset with in total* N_{all} *independent samples over all state-action pairs generated from the nominal environment (for any game* $\mathcal{MG}_i \in \mathcal{M}$*): one has*

$$\begin{aligned} \inf_{\widehat{\pi}} \max_{\mathcal{MG}_i\in\mathcal{M}} \big\{\mathbb{P}_{\mathcal{MG}_i}\big(\mathsf{gap}_{\mathsf{NE}}(\widehat{\pi}) > \varepsilon\big)\big\} &\ge \frac{1}{8}, \\ \inf_{\widehat{\pi}} \max_{\mathcal{MG}_i\in\mathcal{M}} \big\{\mathbb{P}_{\mathcal{MG}_i}\big(\mathsf{gap}_{\mathsf{CCE}}(\widehat{\pi}) > \varepsilon\big)\big\} &\ge \frac{1}{8}, \\ \inf_{\widehat{\pi}} \max_{\mathcal{MG}_i\in\mathcal{M}} \big\{\mathbb{P}_{\mathcal{MG}_i}\big(\mathsf{gap}_{\mathsf{CE}}(\widehat{\pi}) > \varepsilon\big)\big\} &\ge \frac{1}{8}, \end{aligned} \tag{26}$$

provided that

$$N_{\mathsf{all}} \le \frac{C_2 S H^3 \max_{1\le i\le n} A_i}{\varepsilon^2} \min\Big\{H, \frac{1}{\min_{1\le i\le n}\sigma_i}\Big\}. \tag{27}$$

Here, C_2 *is some small enough constant, the infimum is taken over all estimators* $\widehat{\pi}$*, and* $\mathbb{P}_{\mathcal{MG}_i}$ *denotes the probability when the game is* $\mathcal{MG}_i$ *for all* $\mathcal{MG}_i \in \mathcal{M}$*.*

We now highlight several key implications and comparisons that follow from the above results.

Near-optimal sample complexity for RMGs. Theorem 4.1 shows that the proposed model-based algorithm DR-NVI can achieve any robust solution among ε-robust {NE, CCE, CE} when the total number of samples exceeds the order of

$$\widetilde{O}\bigg(\frac{SH^3\prod_{1\le i\le n} A_i}{\varepsilon^2} \min\Big\{H, \frac{1}{\min_{1\le i\le n}\sigma_i}\Big\}\bigg). \tag{28}$$

Combining this with the lower bound in (27) of Theorem 4.2 confirms that the sample complexity of DR-NVI is optimal with respect to many salient factors, including $\varepsilon, S, H, \{\sigma_i\}_{1\le i\le n}$. To the best of our knowledge, this is the first near-optimal sample complexity upper bound for solving robust MGs. As illustrated in Figure 2, it uncovers that the sample requirement of DR-NVI depends on all agents' uncertainty levels $\{\sigma_i\}$ and is inversely proportional to $\min_{i\in[n]}\sigma_i$ when $\min_{i\in[n]}\sigma_i \gtrsim 1/H$. Furthermore, the sample complexity of DR-NVI (Theorem 4.1) significantly improve upon the prior art $\widetilde{O}\big(S^4\big(\prod_{i=1}^n A_i\big)^3 H^4/\varepsilon^2\big)$ (Blanchet et al., 2023).

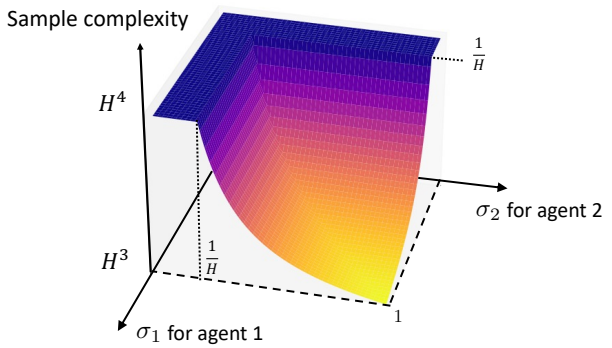


Figure 2. Illustration of the sample complexity of DR-NVI with respect to the uncertainty levels σ_1 and σ_2 for two-player RMGs, where we only highlight the dependency with respect to the horizon length H.

Minimax-optimal sample complexity for single-agent RMDP. We observe that when the size of the action space reduces to one except one agent, i.e. $A_2 = A_3 = \cdots = A_n = 1$, the robust MG simplifies to a single-agent robust Markov decision process (known as RMDP) (Iyengar, 2005). Consequently, the upper bound of (cf. (28)) indicates that a simplified DR-NVI learns an ε-optimal policy for the RMDP associated with the first agent as soon as the sample complexity is on the order of

$$\widetilde{O}\bigg(\frac{SA_1H^3}{\varepsilon^2} \min\Big\{H, \frac{1}{\sigma_1}\Big\}\bigg), \tag{29}$$

which is minimax-optimal in view of the lower bound (cf. (27) of Theorem 4.2). To the best of our knowledge, these findings introduce the first minimax-optimal sample complexity for RMDPs in the finite-horizon setting, complementary to the infinite-horizon result established in Shi et al. (2023).

Benchmarking with standard MGs under non-adaptive sampling. Note that DR-NVI is based on a non-adaptive sampling mechanism from the generative model. Focusing on the same sampling mechanism, we compare the sample complexity of DR-NVI for solving robust MGs with the state-of-the-art approach (model-based NVI) (Zhang et al., 2020a; Liu et al., 2021) for solving standard MGs as below[3]:

$$\text{Standard MGs (by NVI):} \quad \widetilde{O}\bigg(\frac{S\prod_{i=1}^n A_i H^4}{\varepsilon^2}\bigg)$$

[3] Zhang et al. (2020a) considered a two-player zero-sum standard MGs in the infinite-horizon setting. Liu et al. (2021) considered both two-player zero-sum and multi-player general sum standard MGs in online setting. We show the best possible outcomes after transferring into our settings

Robust MGs (by our DR-NVI in Theorem 4.1):

$$\begin{cases} \widetilde{O}\Big(\frac{S\prod_{i=1}^{n} A_i H^4}{\varepsilon^2}\Big) & \text{if } 0 < \min\limits_{1\le i\le n} \sigma_i \lesssim \frac{1}{H} \\ \widetilde{O}\Big(\frac{S\prod_{i=1}^{n} A_i H^3}{\varepsilon^2 \min_{1\le i\le n}\sigma_i}\Big) & \text{if } \frac{1}{H} \lesssim \min\limits_{1\le i\le n} \sigma_i < 1 \end{cases}. \quad (30)$$

It shows that DR-NVI achieves enhanced robustness against model uncertainty in comparison to the prior art NVI for standard MGs, using the same or even sometimes fewer number of samples ($\min_{1\le i\le n}\sigma_i \gtrsim 1/H$). In particular, as illustrated in Figure 2,

- *When* $0 < \min_{1\le i\le n}\sigma_i \lesssim \frac{1}{H}$: the sample complexity dependency of DR-NVI on H matches that of NVI in the order of H^4.
- *When* $\min_{1\le i\le n}\sigma_i \gtrsim \frac{1}{H}$: DR-NVI's sample complexity decreases towards H^3 as $\min_{1\le i\le n}\sigma_i$ increases, which improves upon the sample complexity of NVI for standard MGs by a factor of $H\min_{1\le i\le n}\sigma_i$ that goes to H when $\min_{1\le i\le n}\sigma_i = O(1)$.

Technical challenges and insights. Compared to robust single-agent RL, robust MARL introduces complex statistical dependencies due to game-theoretical interactions between multiple agents and their natural adversaries to choose the worst-case transitions for each agent. Additionally, robust MGs are more intricate than standard MGs since the agents' payoffs become highly nonlinear without closed form, in contrast to being linear in standard MGs. To mitigate these challenges, we carefully control the statistical errors and exploit technical tools from distributionally robust optimization to achieve a near-optimal upper bound. Additionally, note that the established lower bound (Theorem 4.2) is the first information-theoretic lower bound for solving robust MGs, which is achieved by creating a new class of hard instances for the tightness with respect to H and uncertainty levels $\{\sigma_i\}_{1\le i\le n}$.

5. Conclusion

Providing robustness guarantees is a pressing need for RL, one that is especially crucial in multi-agent RL (MARL) since game-theoretical interactions between agents bring in extra instability. We address the vulnerability of MARL to environmental uncertainty by focusing on robust Markov games (RMGs) that consider robustness against worst-case distribution shifts of the shared environment. We design a provable sample-efficient model-based algorithm (DR-NVI) with a finite-sample complexity guarantee. In addition, we provide a lower bound for solving RMGs, which highlights that DR-NVI has near-optimal sample complexity with respect to the size of the state space, the target accuracy, and the horizon length. To the best of our knowledge, this is the first algorithm with near-optimal sample complexity for RMGs. Our work opens up interesting future directions for robust MARL including but not limited to taming the curse of multi-agents and studying other divergence functions for the uncertainty set.

Acknowledgements

The work of L. Shi is supported in part by the Resnick Institute and Computing, Data, and Society Postdoctoral Fellowship at California Institute of Technology. The work of Y. Chi is supported in part by the grants ONR N00014-19-1-2404 and NSF CCF-2106778. The work of A. Wierman is supported in part from the NSF through CNS-2146814, CPS-2136197, CNS-2106403, NGSDI-2105648. The authors thank Shicong Cen, Gen Li, and Yaru Niu for valuable discussions.

Impact Statement

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

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Aumann, R. J. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society*, pp. 1–18, 1987.

Azar, M. G., Munos, R., and Kappen, H. J. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Badrinath, K. P. and Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pp. 511–520. PMLR, 2021.

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

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

Balaji, B., Mallya, S., Genc, S., Gupta, S., Dirac, L., Khare, V., Roy, G., Sun, T., Tao, Y., Townsend, B., et al. Deepracer: Educational autonomous racing platform for experimentation with sim2real reinforcement learning. *arXiv preprint arXiv:1911.01562*, 2019.

Beck, C. L. and Srikant, R. Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12): 1203–1208, 2012.

Bertsimas, D., Gupta, V., and Kallus, N. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*, 2023.

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. 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.

Busoniu, L., Babuska, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Chen, Z., Zhou, D., and Gu, Q. Almost optimal algorithms for two-player zero-sum linear mixture Markov games. In *International Conference on Algorithmic Learning Theory*, pp. 227–261. PMLR, 2022.

Clavier, P., Pennec, E. L., and Geist, M. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

Cui, Q. and Du, S. S. When is offline two-player zero-sum Markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022a.

Cui, Q. and Du, S. S. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *arXiv preprint arXiv:2206.00159*, 2022b.

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

Daskalakis, C., Goldberg, P. W., and Papadimitriou, C. H. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Daskalakis, C., Golowich, N., and Zhang, K. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

Derman, E. and Mannor, S. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

Derman, E., Mankowitz, D. J., Mann, T. A., and Mannor, S. Soft-robust actor-critic policy-gradient. *arXiv preprint arXiv:1803.04848*, 2018.

Dong, J., Li, J., Wang, B., and Zhang, J. Online policy optimization for robust MDP. *arXiv preprint arXiv:2209.13841*, 2022.

Dong, K., Wang, Y., Chen, X., and Wang, L. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*, 2019.

Dou, Z., Yang, Z., Wang, Z., and Du, S. Gap-dependent bounds for two-player markov games. In *International Conference on Artificial Intelligence and Statistics*, pp. 432–455, 2022.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

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

Even-Dar, E. and Mansour, Y. Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25, 2003.

Fang, F., Stone, P., and Tambe, M. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *IJCAI*, pp. 2589–2595, 2015.

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

Gao, R. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

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

Goyal, V. and Grand-Clement, J. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

Han, S., Su, S., He, S., Han, S., Yang, H., and Miao, F. What is the solution for state adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*, 2022.

He, S., Han, S., Su, S., Han, S., Zou, S., and Miao, F. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023.

Ho, C. P., Petrik, M., and Wiesemann, W. Fast bellman updates for robust MDPs. In *International Conference on Machine Learning*, pp. 1979–1988. PMLR, 2018.

Ho, C. P., Petrik, M., and Wiesemann, W. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

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

Jafarnia-Jahromi, M., Wei, C.-Y., Jain, R., and Luo, H. A model-free learning algorithm for infinite-horizon average-reward MDPs with near-optimal regret. *arXiv preprint arXiv:2006.04354*, 2020.

Jia, Z., Yang, L. F., and Wang, M. Feature-based Q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning–a simple, efficient, decentralized algorithm for multiagent RL. *arXiv preprint arXiv:2110.14555*, 2021a.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096, 2021b.

Kakade, S. *On the sample complexity of reinforcement learning*. PhD thesis, University of London, 2003.

Kannan, S. S., Venkatesh, V. L., and Min, B.-C. Smart-llm: Smart multi-agent robot task planning using large language models. *arXiv preprint arXiv:2309.10062*, 2023.

Kardeş, E., Ordóñez, F., and Hall, R. W. Discounted robust stochastic games and an application to queueing control. *Operations research*, 59(2):365–382, 2011.

Kaufman, D. L. and Schaefer, A. J. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

Kearns, M., Mansour, Y., and Ng, A. Y. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49(2-3):193–208, 2002.

Kearns, M. J. and Singh, S. P. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pp. 996–1002, 1999.

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*, 2020.

Kumar, N., Derman, E., Geist, M., Levy, K., and Mannor, S. Policy gradient for s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13589*, 2023.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Lee, J., Jeon, W., Lee, B., Pineau, J., and Kim, K.-E. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Breaking the sample size barrier in model-based reinforcement learning with a generative model. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Li, G., Shi, L., Chen, Y., Gu, Y., and Chi, Y. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Li, G., Chi, Y., Wei, Y., and Chen, Y. Minimax-optimal multi-agent RL in Markov games with a generative model. *Advances in Neural Information Processing Systems*, 35: 15353–15367, 2022a.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 2023.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., and Russell, S. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4213–4220, 2019.

Li, Y. and Lan, G. First-order policy optimization for robust policy evaluation. *arXiv preprint arXiv:2307.15890*, 2023.

Li, Y., Zhao, T., and Lan, G. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022b.

Liang, Z., Ma, X., Blanchet, J., Zhang, J., and Zhou, Z. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.

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

Littman, M. L. and Szepesvári, C. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, volume 96, pp. 310–318, 1996.

Littman, M. L. et al. Friend-or-foe Q-learning in general-sum games. In *ICML*, volume 1, pp. 322–328, 2001.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Liu, S. and Su, H. γ-regret for non-episodic reinforcement learning. *arXiv:2002.05138*, 2020.

Liu, Z. and Xu, P. Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. *arXiv preprint arXiv:2402.15399*, 2024.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. Distributionally robust q-learning. In *International Conference on Machine Learning*, pp. 13623–13643. PMLR, 2022.

Ma, S., Chen, Z., Zou, S., and Zhou, Y. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24 (371):1–40, 2023.

Ma, X., Liang, Z., Blanchet, J., Liu, M., Xia, L., Zhang, J., Zhao, Q., and Zhou, Z. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Mankowitz, D. J., Levine, N., Jeong, R., Shi, Y., Kay, J., Abdolmaleki, A., Springenberg, J. T., Mann, T., Hester, T., and Riedmiller, M. Robust reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:1906.07516*, 2019.

Mao, W. and Başar, T. Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games and Applications*, pp. 1–22, 2022.

Moulin, H. and Vial, J.-P. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3):201–221, 1978.

Nash, J. Non-cooperative games. *Annals of mathematics*, pp. 286–295, 1951.

Nilim, A. and El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

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

Pan, Y., Chen, Y., and Lin, F. Adjustable robust reinforcement learning for online 3d bin packing. *arXiv preprint arXiv:2310.04323*, 2023.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35: 32211–32224, 2022.

Pananjady, A. and Wainwright, M. J. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67 (1):566–585, 2020.

Rahimian, H. and Mehrotra, S. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Ramesh, S. S., Sessa, P. G., Hu, Y., Krause, A., and Bogunovic, I. Distributionally robust model-based reinforcement learning with large state spaces. *arXiv preprint arXiv:2309.02236*, 2023.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Roughgarden, T. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

Roy, A., Xu, H., and Pokutta, S. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

Saloner, G. Modeling, game theory, and strategic management. *Strategic management journal*, 12(S2):119–136, 1991.

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

Shi, L. and Chi, Y. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 19967–20025. PMLR, 2022.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018.

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

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

Slumbers, O., Mguni, D. H., Blumberg, S. B., Mcaleer, S. M., Yang, Y., and Wang, J. A game-theoretic framework for managing risk in multi-agent systems. In *International Conference on Machine Learning*, pp. 32059–32087. PMLR, 2023.

Smirnova, E., Dohmatob, E., and Mary, J. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

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

Tamar, A., Mannor, S., and Xu, H. Scaling up robust MDPs using function approximation. In *International conference on machine learning*, pp. 181–189. PMLR, 2014.

Tian, Y., Wang, Y., Yu, T., and Sra, S. Online learning in unknown markov games. In *International conference on machine learning*, pp. 10279–10288. PMLR, 2021.

Tsybakov, A. B. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

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

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

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

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wainwright, M. J. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

Wang, H., Shi, L., and Chi, Y. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. A finite sample complexity bound for distributionally robust Q-learning. *arXiv preprint arXiv:2302.13203*, 2023a.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*, 2023b.

Wang, S., Si, N., Blanchet, J., and Zhou, Z. Sample complexity of variance-reduced distributionally robust Q-learning. *arXiv preprint arXiv:2305.18420*, 2023c.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34, 2021.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on Learning Theory*, pp. 4259–4299. PMLR, 2021.

Wiesemann, W., Kuhn, D., and Rustem, B. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Wolff, E. M., Topcu, U., and Murray, R. M. Robust control of uncertain markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3372–3379. IEEE, 2012.

Woo, J., Joshi, G., and Chi, Y. The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *arXiv preprint arXiv:2305.10697*, 2023.

Woo, J., Shi, L., Joshi, G., and Chi, Y. Federated offline reinforcement learning: Collaborative single-policy coverage suffices. *arXiv preprint arXiv:2402.05876*, 2024.

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

Xu, H. and Mannor, S. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

Xu, Z., Panaganti, K., and Kalathil, D. Improved sample complexity bounds for distributionally robust reinforcement learning. *arXiv preprint arXiv:2303.02783*, 2023.

Yan, Y., Li, G., Chen, Y., and Fan, J. The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*, 2022a.

Yan, Y., Li, G., Chen, Y., and Fan, J. Model-based reinforcement learning is minimax-optimal for offline zero-sum markov games. *arXiv preprint arXiv:2206.04044*, 2022b.

Yang, K., Yang, L., and Du, S. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Yang, L. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004, 2019.

Yang, W., Zhang, L., and Zhang, Z. Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. Avoiding model estimation in robust markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*, 2023.

Yang, Y. and Ma, C. $o(t^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. *arXiv preprint arXiv:2209.12430*, 2022.

Yeh, C., Meng, C., Wang, S., Driscoll, A., Rozi, E., Liu, P., Lee, J., Burke, M., Lobell, D. B., and Ermon, S. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021.

Yin, M. and Wang, Y.-X. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

Zanette, A., Kochenderfer, M. J., and Brunskill, E. Almost horizon-free structure-aware best policy identification with a generative model. *Advances in Neural Information Processing Systems*, 32, 2019.

Zeng, L., Qiu, D., and Sun, M. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. *Applied Energy*, 324: 119688, 2022.

Zhang, H., Chen, H., Boning, D., and Hsieh, C.-J. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021a.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33, 2020a.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020b.

Zhang, K., Sun, T., Tao, Y., Genc, S., Mallya, S., and Basar, T. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33:10571–10583, 2020c.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021b.

Zhang, R., Hu, Y., and Li, N. Regularized robust mdps and risk-sensitive mdps: Equivalence, policy gradient, and sample complexity. *arXiv preprint arXiv:2306.11626*, 2023a.

Zhang, Z., Ji, X., and Du, S. S. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020d.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *arXiv preprint arXiv:2006.03864*, 2020e.

Zhang, Z., Chen, Y., Lee, J. D., and Du, S. S. Settling the sample complexity of online reinforcement learning. *arXiv preprint arXiv:2307.13586*, 2023b.

Zhang, Z., Sun, Y., Huang, F., and Miao, F. Safe and robust multi-agent reinforcement learning for connected autonomous vehicles under state perturbations. *arXiv preprint arXiv:2309.11057*, 2023c.

Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z., and Yang, Z. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. *arXiv preprint arXiv:2202.07511*, 2022.

Zhou, M., Luo, J., Villella, J., Yang, Y., Rusu, D., Miao, J., Zhang, W., Alban, M., Fadakar, I., Chen, Z., et al. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. *arXiv preprint arXiv:2010.09776*, 2020.

Zhou, Z. and Liu, G. Robustness testing for multi-agent reinforcement learning: State perturbations on critical agents. *arXiv preprint arXiv:2306.06136*, 2023.

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

A. Related Works

In this section, we discuss a non-exhaustive set of related works, limiting our discussions primarily to provable RL algorithms in the tabular setting, which are most related to this paper.

Finite-sample studies of standard Markov games. Multi-agent reinforcement learning (MARL), originated from the seminal work (Littman, 1994), has been widely studied under the framework of standard Markov games (Shapley, 1953); see Busoniu et al. (2008); Zhang et al. (2021b); Oroojlooy & Hajinezhad (2023) for detailed reviews. There has been no shortage of provably convergent MARL algorithms with asymptotic guarantees (Littman & Szepesvári, 1996; Littman et al., 2001).

A line of recent efforts have concentrated on understanding and developing algorithms for standard MGs with non-asymptotic guarantees (finite-sample analysis). Within this field, Nash equilibrium (NE) is arguably one of the most compelling solution concepts for standard MGs. Research on calculating NE primarily focuses on an important basic class: standard two-player zero-sum MGs (Bai & Jin, 2020; Chen et al., 2022; Mao & Başar, 2022; Wei et al., 2017; Tian et al., 2021; Cui & Du, 2022a;b; Zhong et al., 2022; Jia et al., 2019; Yang & Ma, 2022; Yan et al., 2022b; Dou et al., 2022; Wei et al., 2021). This focus arises because computing NEs in scenarios beyond the standard two-player zero-sum MGs is generally computationally intractable (i.e., PPAD-complete) (Daskalakis, 2013; Daskalakis et al., 2009). For discounted infinite-horizon two-player zero-sum Markov games, the state-of-the-art sample complexity for learning NE (Zhang et al., 2020e) remains suboptimal due to the "curse of multiple agents" issue (Zhang et al., 2020e). In contrast, for episodic finite-horizon two-player zero-sum Markov games standard MGs, Bai et al. (2020); Jin et al. (2021a); Li et al. (2022a) have overcome this curse, progressively achieving minimax-optimal sample complexity in the order of $O(S\max_{1\leq i\leq n} A_i H^4/\varepsilon^2)$. Besides NE, Jin et al. (2021a); Daskalakis et al. (2022); Mao & Başar (2022); Song et al. (2021); Li et al. (2022a); Liu et al. (2021) have extended this achievement to other computationally tractable solution concepts (e.g., CE/CCE) in general-sum multi-player MGs. Focusing on the same non-adaptive sampling mechanism considered in this work, the sample complexity for learning NE/CE/CCE in standard MGs with the state-of-the-art approaches (Zhang et al., 2020e; Liu et al., 2021) still suffers from the curse of multiple agents, calculated as $O(S\prod_{1\leq i\leq n} A_i H^4/\varepsilon^2)$.

Robustness in MARL. Despite significant advances in standard MARL, current algorithms may fail dramatically due to perturbations or uncertainties in game components, resulting in significantly deviated equilibrium, as illustrated in Figure 1. A growing body of research is now addressing the robustness of MARL algorithms against uncertainties in various components of Markov games, such as state (Han et al., 2022; He et al., 2023; Zhou & Liu, 2023; Zhang et al., 2023c), environment (reward and transition kernel), the type of agents (Zhang et al., 2021a), or other agents' policies (Li et al., 2019; Kannan et al., 2023); see Vial et al. (2022) for a recent review.

This work considers the robustness against environmental uncertainty, adopting distributionally robust optimization (DRO) that has primarily been investigated in the context of supervised learning (Rahimian & Mehrotra, 2019; Gao, 2020; Bertsimas et al., 2018; Duchi & Namkoong, 2018; Blanchet & Murthy, 2019). Applying DRO for single-agent RL (Iyengar, 2005) to handle model uncertainty has garnered significant attention. When turning to MARL, the problem is conceptualized as robust Markov games within the DRO framework, an area that remains relatively underexplored with only a few provable algorithms developed (Zhang et al., 2020c; Kardeş et al., 2011; Ma et al., 2023; Blanchet et al., 2023). Notably, Kardeş et al. (2011) verifies the existence of Nash equilibrium for robust Markov games under mild assumptions; Zhang et al. (2020c) derives asymptotic convergence for a Q-learning type algorithm under certain conditions; Ma et al. (2023); Blanchet et al. (2023) are the most related works that provide algorithms with finite-sample guarantees for various types of uncertainty set. Especially, Ma et al. (2023) considers a restricted uncertainty level that could fail to bring robustness to MARL in certain scenarios. In particular, as the required accuracy level (ε goes to zero or the robust MGs has a small minimal positive transition probabilities ($p_{\min}\to 0$), the required uncertainty level becomes quite restrictive (obeying $\sigma_i \leq \max\{\frac{\varepsilon}{SH^2}, \frac{p_{\min}}{H}\}$ for all $i\in[n]$) — potentially reducing robust MARL to standard MARL and failing to maintain desired robustness.

Single-agent distributionally robust RL (robust MDPs). For single-agent RL, considering robustness to model uncertainty using DRO framework — i.e., distributionally robust dynamic programming and robust MDPs — has gained significant attention across both theoretical and practical domains (Iyengar, 2005; Xu & Mannor, 2012; Wolff et al., 2012; Kaufman & Schaefer, 2013; Ho et al., 2018; Smirnova et al., 2019; Ho et al., 2021; Goyal & Grand-Clement, 2022; Derman & Mannor, 2020; Tamar et al., 2014; Badrinath & Kalathil, 2021; Roy et al., 2017; Derman et al., 2018; Mankowitz et al., 2019). Recently, a substantial body of work has been dedicated to exploring the finite-sample performance of provable

robust single-agent RL algorithms, where different sampling mechanisms, diverse divergence function of the uncertainty set, and other related problems/issues has been investigated a lot (Yang et al., 2022; Panaganti & Kalathil, 2022; Zhou et al., 2021; Shi & Chi, 2022; Wang et al., 2023a; Blanchet et al., 2023; Liu et al., 2022; Wang et al., 2023c; Liang et al., 2023; Shi et al., 2023; Wang & Zou, 2021; Xu et al., 2023; Dong et al., 2022; Badrinath & Kalathil, 2021; Ramesh et al., 2023; Panaganti et al., 2022; Ma et al., 2022; Wang et al., 2023b; Li et al., 2022b; Kumar et al., 2023; Clavier et al., 2023; Yang et al., 2023; Zhang et al., 2023a; Li & Lan, 2023; Wang et al., 2024).

Among the studies of robust MDPs, those particularly relevant to this paper employ the uncertainty set using total variation (TV) distance in a tabular setting (Yang et al., 2022; Panaganti & Kalathil, 2022; Xu et al., 2023; Dong et al., 2022; Liu & Xu, 2024). It has been established that solving robust MDPs requires no more samples than solving standard MDPs in terms of the sample requirement (Shi et al., 2023) with a generative model. However, robust MARL involves additional complexities compared to robust single-agent RL. It remains an open question whether the findings from robust MDPs can be generalized to robust MARL, which includes more technical challenges and strategic interactions. Our work takes a step towards the question, confirming that similar phenomena apply in robust MARL, albeit with increased difficulties due to the multi-agent dynamics.

RL with a generative model. Access to a generative model (or simulator) serves as a fundamental and idealistic sampling protocol that has been widely used to study finite-sample guarantees for diverse types of RL algorithms, such as various model-based, model-free, and policy-based algorithms (Kearns et al., 2002; Agarwal et al., 2020; Azar et al., 2013; Li et al., 2020; Sidford et al., 2018; Wainwright, 2019; Li et al., 2023; Kakade, 2003; Pananjady & Wainwright, 2020; Khamaru et al., 2020; Even-Dar & Mansour, 2003; Beck & Srikant, 2012; Zanette et al., 2019; Yang & Wang, 2019; Woo et al., 2023). This work follows this fundamental protocol with a non-adaptive sampling mechanism to understand and design algorithms for robust Markov games. Besides generative model, there also exist other sampling protocols that involve more realistic scenarios such as online exploration setting (Dong et al., 2019; Zhang et al., 2020d;e; Jafarnia-Jahromi et al., 2020; Liu & Su, 2020; Yang et al., 2021; Zhang et al., 2023b; Li et al., 2021) or offline setting (Xie et al., 2021; Rashidinejad et al., 2021; Jin et al., 2021b; Yin & Wang, 2021; Yan et al., 2022a; Uehara & Sun, 2021; Woo et al., 2024; Shi et al., 2022; Li et al., 2024), which are interesting directions in the future.

B. Preliminaries

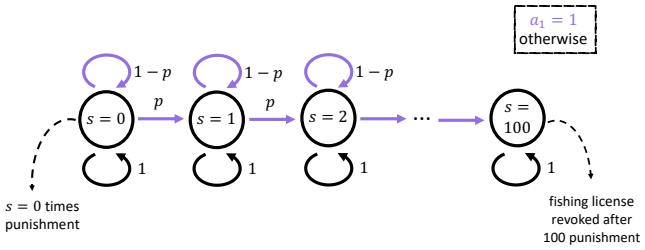
B.1. Details of the example shown in Figure 1

The standard Markov game for fishing protection. To simulate a scenario of defense against illegal fishing, we can formulate a two-player general sum finite-horizon standard Markov game between a fisher (the first player) and a police officer (the second player). This MG can be represented as $\mathcal{MG}^e = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le 2}, p, r, H\big\}$. Here, $\mathcal{S} := \{0, 1, \cdots, 100\}$ is the state space, where each state $s \in \mathcal{S}$ represents the number of punishments received by the fisherman, with the license being revoked at $s = 100$; $\mathcal{A}_1 = \mathcal{A}_2 = \{0, 1\}$ is the action space. At each time step (round), the fisher chooses a_1 among space $\mathcal{A}_1 = \{$legal fishing (0), illegal fishing (1)$\}$, while the officer chooses a_2 among $\mathcal{A}_2 = \{$no patrols (0), go patrols (1)$\}$; H is the horizon-length; the transition kernel is governed by a model parameter $p \in [0, 1]$, shown in Figure 3(a) (a detailed version of Figure 1(a)), specified as

$$\forall h \in [H]: \quad P_h(s' \,|\, s, a_1, a_2) = \begin{cases} p\mathbb{1}(s' = s+1) + (1-p)\mathbb{1}(s' = s) & \text{if} \quad s \in \mathcal{S} \setminus \{100\}, a_1 = 1, \\ \mathbb{1}(s' = s) & \text{otherwise.} \end{cases} \tag{31}$$

In words, the state s transit to $s' = s + 1$ with probability p when $a_1 = 1$, otherwise staying in $s' = s$, i.e., In addition, $r = \{r_{i,h}\}_{i\in\{1,2\},h\in[H]}$ represents the immediate reward (benefit) function of two players at each time step $h \in [H]$. Here, we consider time-invariant reward function $r_{i,h} = r_i$ for all $h \in [H]$. In particular, at any time step $h \in [H]$, $r_1(s, a_1, a_2, s')$ (resp. $r_2(s, a_1, a_2, s')$) denotes the immediate benefit that the first agent (resp. the second player) receives conditioned on the current state s, the actions of two players (a_1, a_2), and the next state s'. The reward function for any state $s \in \mathcal{S} \setminus \{100\}$ is defined in Figure 3(b). And the reward function at state $s = 100$ for two players is specified as below:

$$\begin{aligned} \forall a_2 \in \{0, 1\}: \quad & r_1(100, 0, a_2, 100) = -1 \quad \text{and} \quad r_1(100, 1, a_2, 100) = -20p \\ & r_2(100, 0, 0, 100) = 1 \quad \text{and} \quad r_2(100, 0, 1, 100) = 0 \\ & r_2(100, 1, 0, 100) = 1 \quad \text{and} \quad r_2(100, 1, 1, 100) = 3 - 2p. \end{aligned} \tag{32}$$



(a) States and Transitions

		Agent 2: Officer	
		$a_2 = 0$ [no patrols]	$a_2 = 1$ [go patrols]
Agent 1: Fisher	$a_1 = 0$ legal fishing	$(-1, 1)$	$(-1, 0)$
	$a_1 = 1$ illegal fishing	$\begin{cases}(-20,1) & \text{if } s' = s+1 \\ (0,1) & \text{if } s' = s\end{cases}$	$\begin{cases}(-20,1) & \text{if } s' = s = 1 \\ (0,3) & \text{if } s' = s\end{cases}$

(b) Reward: $(r_1(s, a_1, a_2, s'), r_2(s, a_1, a_2, s'))$ for $s = \{0, 1, \cdots, 99\}$

Figure 3. (a) shows the transition kernels of the game at each time step h. (b) illustrates the immediate reward function of two agents.

Computing the Nash equilibrium (NE). Notice that the NE of a standard Markov game is indeed a series of NE of the matrix games associated with the Q-function at each time step $h \in [H]$. We denote the NE of $\mathcal{MG}^e$ as $\pi^\star = (\mu^\star, \nu^\star) = \{\mu_h^\star, \nu_h^\star\}_{h\in[H]}$ with $\mu_h^\star : \mathcal{S} \mapsto \Delta(\mathcal{A}_1), \nu_h^\star : \mathcal{S} \mapsto \Delta(\mathcal{A}_2)$ for all $h \in [H]$. To proceed, we start from characterizing the Q-function and Bellman consistency equation of the fishing protection game.

It is easily verified that for time step $H+1$, one has for any joint policy $\pi = (\mu, \nu)$,

$$\forall (i, a_1, a_2) \in \{1, 2\} \times \mathcal{A}_1 \times \mathcal{A}_2 : \quad Q_{i,H+1}^{\pi,P}(s, a_1, a_2) = Q_{i,H+1}^{\pi,P}(s', a_1, a_2) = 0. \tag{33}$$

Then, we characterize the Bellman consistency equation at time step $h = H, H-1, \cdots, 1$ for the optimal policy $\pi^\star$. Notice that the rewards and the transition kernels have similar structures for all states except $s = 100$. So we start from the cases when $s \in \mathcal{S} \setminus \{100\}$. Recalling the definition of Q-function in (3), the reward function r (defined in Figure 3(b)) and the transition kernel in (31), we have for any state $s \in \mathcal{S} \setminus \{100\}$ and any time step $h \in [H]$, the Q-function of the fisher (the first player) obeys

$$\begin{aligned}
Q_{1,h}^{\pi^\star,P}(s, 0, 0) &= -1 + V_{1,h+1}^{\pi^\star,P}(s) \\
Q_{1,h}^{\pi^\star,P}(s, 0, 1) &= -1 + V_{1,h+1}^{\pi^\star,P}(s) \\
Q_{1,h}^{\pi^\star,P}(s, 1, 0) &= -20p + pV_{1,h+1}^{\pi^\star,P}(s+1) + (1-p)V_{1,h+1}^{\pi^\star,P}(s) \\
Q_{1,h}^{\pi^\star,P}(s, 1, 1) &= -20p + pV_{1,h+1}^{\pi^\star,P}(s+1) + (1-p)V_{1,h+1}^{\pi^\star,P}(s).
\end{aligned} \tag{34}$$

Similarly, for the officer (the second player), we observe that for any state $s \in \mathcal{S} \setminus \{100\}$ and any time step $h \in [H]$:

$$\begin{aligned}
Q_{2,h}^{\pi^\star,P}(s, 0, 0) &= 1 + V_{2,h+1}^{\pi^\star,P}(s) \\
Q_{2,h}^{\pi^\star,P}(s, 0, 1) &= 0 + V_{2,h+1}^{\pi^\star,P}(s), \\
Q_{2,h}^{\pi^\star,P}(s, 1, 0) &= 1 + pV_{2,h+1}^{\pi^\star,P}(s+1) + (1-p)V_{2,h+1}^{\pi^\star,P}(s), \\
Q_{2,h}^{\pi^\star,P}(s, 1, 1) &= 3 - 2p + pV_{2,h+1}^{\pi^\star,P}(s+1) + (1-p)V_{2,h+1}^{\pi^\star,P}(s).
\end{aligned} \tag{35}$$

Armed with above results, we are now ready to show that the NE for all $(h, s) \in [H] \times \mathcal{S}$ are the same, which determined by the model parameter p as below:

$$\forall (h, s) \in [H] \times \mathcal{S} : \quad \begin{cases} \pi_h^\star(s) = \pi_B = (0, 0) & \text{if } p > 0.05 \\ \pi_h^\star(s) = \pi_A = (1, 1) & \text{if } p \le 0.05. \end{cases} \tag{36}$$

We will verify it by induction as below:

- *Base case: when $h = H$.* Applying (34) and (35) for $h = H$ with the fact in (33), we arrive at for any state $s \in \mathcal{S} \setminus \{100\}$:

$$\forall a_2 \in \{0, 1\} : \quad Q_{1,H}^{\pi^\star,P}(s, 0, a_2) = -1 \quad \text{and} \quad Q_{1,H}^{\pi^\star,P}(s, 1, a_2) = -20p$$

$$Q_{2,H}^{\pi^\star,P}(s,0,0)=1 \quad \text{and} \quad Q_{2,H}^{\pi^\star,P}(s,0,1)=0$$
$$Q_{2,H}^{\pi^\star,P}(s,1,0)=1 \quad \text{and} \quad Q_{2,H}^{\pi^\star,P}(s,1,1)=3-2p \tag{37}$$

Similarly, when state $s=100$, recalling the reward function in (32), we achieve the same Q-function on state $s=100$. Therefore, one has for all $s\in\mathcal{S}$:

$$\begin{aligned}\forall a_2\in\{0,1\}: \quad & Q_{1,H}^{\pi^\star,P}(s,0,a_2)=-1 \quad \text{and} \quad Q_{1,H}^{\pi^\star,P}(s,1,a_2)=-20p\\ & Q_{2,H}^{\pi^\star,P}(s,0,0)=1 \quad \text{and} \quad Q_{2,H}^{\pi^\star,P}(s,0,1)=0\\ & Q_{2,H}^{\pi^\star,P}(s,1,0)=1 \quad \text{and} \quad Q_{2,H}^{\pi^\star,P}(s,1,1)=3-2p.\end{aligned} \tag{38}$$

Consequently, in view of (44), it can be verified that if $p<0.05$ (resp. $p>0.05$), the unique NE of two agents on any state $s\in\mathcal{S}$ at time step H is the policy pair $\pi_H^\star(s)=(\mu_H^\star(s),\nu_H^\star(s))=(1,1)$ (resp. $\pi_H^\star(s)=(\mu_H^\star(s),\nu_H^\star(s))=(0,0)$), leading to Nash $\pi_A\coloneqq(1,1)$ when $p=p_A=0.049$ (resp. Nash $\pi_B\coloneqq(0,0)$ when $p=p_B=0.051$).

In addition, we observe the optimal value function satisfies that:

$$\forall s\in\mathcal{S}: \begin{cases} V_{1,H}^{\pi^\star,P}(s)=-1 \quad \text{and} \quad V_{2,H}^{\pi^\star,P}(s)=1 & \text{if } p>0.05\\ V_{1,H}^{\pi^\star,P}(s)=-20p \quad \text{and} \quad V_{2,H}^{\pi^\star,P}(s)=3-2p & \text{if } p\le 0.05\end{cases}. \tag{39}$$

- *Induction.* The rest of this paragraph is to verify (36) for all $(h,s)\in[H-1]\times\mathcal{S}$ by induction. So suppose (36) holds for time step $h+1$, then we will show that it also holds for time step h.

  To begin with, we introduce the following claim which will be verified in Appendix B.1.1: for any policy $\pi=(\mu,\nu)$ and any $s,s'\in\mathcal{S}$:

$$\forall(i,h)\in\{1,2\}\times[H]: \quad V_{i,h}^{\pi,P}(s)=V_{i,h}^{\pi,P}(s'). \tag{40}$$

  To proceed, armed with the fact in (40), invoking the results in (34) and (35) yields that for all $s\in\mathcal{S}$:

$$\begin{aligned} &Q_{1,h}^{\pi^\star,P}(s,0,0)=-1+V_{1,h+1}^{\pi^\star,P}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,P}(s,0,0)=1+V_{2,h+1}^{\pi^\star,P}(s)\\ &Q_{1,h}^{\pi^\star,P}(s,0,1)=-1+V_{1,h+1}^{\pi^\star,P}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,P}(s,0,1)=0+V_{2,h+1}^{\pi^\star,P}(s)\\ &Q_{1,h}^{\pi^\star,P}(s,1,0)=-20p+V_{1,h+1}^{\pi^\star,P}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,P}(s,1,0)=1+V_{2,h+1}^{\pi^\star,P}(s)\\ &Q_{1,h}^{\pi^\star,P}(s,1,1)=-20p+V_{1,h+1}^{\pi^\star,P}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,P}(s,1,1)=3-2p+V_{2,h+1}^{\pi^\star,P}(s).\end{aligned} \tag{41}$$

  The above fact directly indicates that at time step h, the NE of the matrix games associated with the payoff $Q_{1,h}^{\pi^\star,P}(s)$ and $Q_{2,h}^{\pi^\star,P}(s)$ satisfies

$$\forall s\in\mathcal{S}: \quad \begin{cases}\pi_h^\star(s)=(0,0) & \text{if } p>0.05\\ \pi_h^\star(s)=(1,1) & \text{if } p\le 0.05.\end{cases} \tag{42}$$

Summing up the base case and the induction results, we complete the proof for (36).

The robust MG and computing the robust Nash equilibrium (robust NE). When turns to the robust formulation of the fishing protection game, we construct a robust Markov game represented as $\mathcal{MG}_{\mathsf{rob}}^e=\big\{\mathcal{S},\{\mathcal{A}_i\}_{1\le i\le 2},p^0,\sigma,r,H\big\}$, where $\mathcal{S},\{\mathcal{A}_i\}_{1\le i\le 2},r,H$ are the same as those defined in the standard MG $\mathcal{MG}^e$. Note that this example is designed to illustrate general environmental uncertainty (includes both the reward and transition kernel uncertainty) and is not tailored to the specific class of robust MGs defined in Section 3. For simplicity, let each agent consider that the model parameter p can perturb around some nominal one p^0 with uncertainty level $\sigma=0.005$, i,e., $p\in[p^0-\sigma,p^0+\sigma]$. Other components of the transition kernel is not allowed to perturb. With abuse of notation, for any joint policy π, we still denote the robust value function (resp. robust Q-function) for i-th agent at time step h as $V_{i,h}^{\pi,\sigma}$ (resp. $Q_{i,h}^{\pi,\sigma}$). In addition, we denote the robust NE of $\mathcal{MG}_{\mathsf{rob}}^e$ as $\pi^{\star,\sigma}=(\mu^{\star,\sigma},\nu^{\star,\sigma})=\{\mu_h^{\star,\sigma},\nu_h^{\star,\sigma}\}_{h\in[H]}$, where $\mu_h^{\star,\sigma}:\mathcal{S}\mapsto\Delta(\mathcal{A}_1),\nu_h^{\star,\sigma}:\mathcal{S}\mapsto\Delta(\mathcal{A}_2)$.

Observe that in city A (resp. city B), the nominal model parameter $p^0 = 0.049$ (resp. $p^0 = 0.051$). Without loss of generality, we first focus on city A. To proceed, we shall verify the following claim using the same routine for computing NE of the standard MG $\mathcal{MG}^e$ (cf. (36)):

$$\text{In city A}: \quad (\mu_h^{\star,\sigma}(s), \nu_h^{\star,\sigma}(s)) = (0,0), \quad \forall (h,s) \in [H] \times \mathcal{S}. \tag{43}$$

- *Base case: when $h = H$.* Recall the definitions of robust value/Q-function (cf. (11)), one has at time step H: for all $s \in \mathcal{S}$,

$$\begin{aligned} \forall a_2 \in \{0,1\}: \quad & Q_{1,H}^{\pi^\star,\sigma}(s,0,a_2) = -1 \quad \text{and} \quad Q_{1,H}^{\pi^\star,\sigma}(s,1,a_2) = -20(p^0+\sigma) = -1.08 \\ & Q_{2,H}^{\pi^\star,\sigma}(s,0,0) = 1 \quad \text{and} \quad Q_{2,H}^{\pi^\star,\sigma}(s,0,1) = 0 \\ & Q_{2,H}^{\pi^\star,\sigma}(s,1,0) = 1 \quad \text{and} \quad Q_{2,H}^{\pi^\star,\sigma}(s,0,1) = 3 - 2(p^0+\sigma) = 2.892. \end{aligned} \tag{44}$$

  As a result, it is easily verified that the unique robust NE of two agents on any state $s \in \mathcal{S}$ at time step H is the policy pair $(\mu_H^{\star,\sigma}(s), \nu_H^{\star,\sigma}(s)) = (0,0)$.

- *Induction.* First of all, for any policy $\pi = (\mu, \nu)$ and $s, s' \in \mathcal{S}$, similar to (40)

$$\forall (i,h) \in \{1,2\} \times [H]: \quad V_{i,h}^{\pi,\sigma}(s) = V_{i,h}^{\pi,\sigma}(s'). \tag{45}$$

  which indicates that the worst-case performance are indeed influenced by the uncertainty of the reward function but not the transition kernel perturbation. Armed with above fact, invoking the robust Bellman consistency equation, similar to (41), we can achieve that for all $h \in 1, 2, \cdots, H-1$,

$$\begin{aligned} & Q_{1,h}^{\pi^\star,\sigma,\sigma}(s,0,0) = -1 + V_{1,h+1}^{\pi^\star,\sigma,\sigma}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,\sigma,\sigma}(s,0,0) = 1 + V_{2,h+1}^{\pi^\star,\sigma,\sigma}(s) \\ & Q_{1,h}^{\pi^\star,\sigma,\sigma}(s,0,1) = -1 + V_{1,h+1}^{\pi^\star,\sigma,\sigma}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,\sigma,\sigma}(s,0,1) = 0 + V_{2,h+1}^{\pi^\star,\sigma,\sigma}(s) \\ & Q_{1,h}^{\pi^\star,\sigma,\sigma}(s,1,0) = -1.08 + V_{1,h+1}^{\pi^\star,\sigma,\sigma}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,\sigma,\sigma}(s,1,0) = 1 + V_{2,h+1}^{\pi^\star,\sigma,\sigma}(s) \\ & Q_{1,h}^{\pi^\star,\sigma,\sigma}(s,1,1) = -1.08 + V_{1,h+1}^{\pi^\star,\sigma,\sigma}(s) \quad \text{and} \quad Q_{2,h}^{\pi^\star,\sigma,\sigma}(s,1,1) = 2.892 + V_{2,h+1}^{\pi^\star,\sigma,\sigma}(s). \end{aligned} \tag{46}$$

  As a consequence, the robust NE of the matrix games associated with the payoff $Q_{1,h}^{\pi^\star,\sigma,\sigma}(s)$ and $Q_{2,h}^{\pi^\star,\sigma,\sigma}(s)$ satisfies $(\mu_h^{\star,\sigma}(s), \nu_h^{\star,\sigma}(s)) = (0,0)$ for all $h \in 1, 2, \cdots, H-1$.

Summing up the results in the base case and the induction, we verify the unique robust NE for $\mathcal{MG}_{\mathsf{rob}}^e$ in city A as (43). The same unique robust NE can be verified in city B by following the same routine, which we omit for brevity. Thus, we show the unique robust NE in two slightly different environments (city A and city B) are identical.

Deriving the states of executing different equilibrium solutions. In view of (36), we know that the NE of the standard MG $\mathcal{MG}^e$ in city A when $p = p_A = 0.049$ (resp. city B when $p = p_B = 0.051$) is $\pi_A = (1,1)$ (resp. $\pi_B = (0,0)$) for all $(h,s) \in [H] \times \mathcal{S}$. And the MG $\mathcal{MG}^e$ has some one-way transition structure, namely state s can only transit to itself or a larger state $s+1$, while not any states $s' < s$. So as long as H is large enough, the final state of executing $\pi_A = (1,1)$ will be state $s = 100$ with the fishing license revoked since the fisher will always do illegal fishing ($a_1 = 1$). The agents who execute the joint policy $\pi_B = (0,0)$ or the robust NE $(\mu_h^{\star,\sigma}(s), \nu_h^{\star,\sigma}(s)) = (0,0)$ will stay in $s = 0$ with no punishment since the fisher will never choose illegal fishing ($a_1 = 1$).

B.1.1. Proof of claim (40)

We will proof (40) by induction. Note that the base case when $h = H$ has already been verified in (39).

Then suppose the claim holds at time step $h+1$, i.e.,

$$\forall (i,s,s') \in \{1,2\} \times \mathcal{S} \times \mathcal{S}: \quad V_{i,h+1}^{\pi,P}(s) = V_{i,h+1}^{\pi,P}(s'), \tag{47}$$

it remains to show that the claim holds at time step h as well.

Towards this, we first consider the cases when state $s \in \mathcal{S} \setminus \{100\}$. Recall the recursion in (34), we arrive at

$$\begin{aligned}
Q_{1,h}^{\pi,P}(s,0,0) &= -1 + V_{1,h+1}^{\pi,P}(s) \\
Q_{1,h}^{\pi,P}(s,0,1) &= -1 + V_{1,h+1}^{\pi,P}(s) \\
Q_{1,h}^{\pi,P}(s,1,0) &= -20p + pV_{1,h+1}^{\pi,P}(s+1) + (1-p)V_{1,h+1}^{\pi^\star,P}(s) \overset{\text{(i)}}{=} -20p + V_{1,h+1}^{\pi,P}(s) \\
Q_{1,h}^{\pi,P}(s,1,1) &= -20p + pV_{1,h+1}^{\pi,P}(s+1) + (1-p)V_{1,h+1}^{\pi^\star,P}(s) \overset{\text{(ii)}}{=} -20p + V_{1,h+1}^{\pi,P}(s),
\end{aligned} \tag{48}$$

where (i) and (ii) holds by the induction assumption in (50).

Analogously, recalling (35) for the second player (protector), we arrive at for any state $s \in \mathcal{S} \setminus \{100\}$ and time step $h \in [H]$,

$$\begin{aligned}
Q_{2,h}^{\pi,P}(s,0,0) &= 1 + V_{2,h+1}^{\pi,P}(s) \\
Q_{2,h}^{\pi,P}(s,0,1) &= 0 + V_{2,h+1}^{\pi,P}(s), \\
Q_{2,h}^{\pi,P}(s,1,0) &= 1 + V_{2,h+1}^{\pi,P}(s), \\
Q_{2,h}^{\pi,P}(s,1,1) &= 3 - 2p + V_{2,h+1}^{\pi,P}(s).
\end{aligned} \tag{49}$$

Combining (48) and (49) gives that for any $s, s' \in \mathcal{S} \setminus \{100\}$,

$$\forall (i, a_1, a_2) \in \{1,2\} \times \Delta(\mathcal{A}_1) \times \Delta(\mathcal{A}_2): \quad Q_{i,h}^{\pi,P}(s,a_1,a_2) = Q_{i,h}^{\pi,P}(s',a_1,a_2), \tag{50}$$

which indicates

$$V_{i,h}^{\pi,P}(s) = \mathbb{E}_{(a_1,a_2)\in\mu(s)\times\mu(s)}[Q_{i,h}^{\pi,P}(s,a_1,a_2)] = \mathbb{E}_{(a_1,a_2)\in\mu(s)\times\mu(s)}[Q_{i,h}^{\pi,P}(s',a_1,a_2)] = V_{i,h}^{\pi,P}(s'). \tag{51}$$

Similarly, when $s = 100$, it can be verified that (48) and (49) also hold. Therefore, we complete the induction argument by observing that for all $s, s' \in \mathcal{S}$, $V_{i,h}^{\pi,P}(s) = V_{i,h}^{\pi,P}(s')$ is satisfied.

B.2. Additional notation and basic facts

For convenience, for any two vectors $x = [x_i]_{1\le i\le n}$ and $y = [y_i]_{1\le i\le n}$, the notation $x \le y$ (resp. $x \ge y$) means $x_i \le y_i$ (resp. $x_i \ge y_i$) for all $1 \le i \le n$. We denote by $x \circ y = \big[x(s)\cdot y(s)\big]_{s\in\mathcal{S}}$ the Hadamard product of any two vectors $x, y \in \mathbb{R}^S$. And for any vecvor x, we let $x^{\circ 2} = \big[x(s,a)^2\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ (resp. $x^{\circ 2} = \big[x(s)^2\big]_{s\in\mathcal{S}}$). With slight abuse of notation, we denote 0 (resp. 1) as the all-zero (resp. all-one) vector, and $e_i \in \mathbb{R}^S$ as a S-dimensional basis vector with the i-th entry being 1 and others being 0. Recall that we abbreviate the subscript ρ_{TV} when the divergence function is specified to TV distance to write $\mathcal{U}^\sigma(\cdot) = \mathcal{U}^\sigma_{\rho_{\mathsf{TV}}}(\cdot)$.

Additional matrix notation. For any $(i,h) \in [n] \times [H]$, we recall or introduce some additional notation and matrix notation that is useful throughout the analysis

- $r_{i,h} = [r_{i,h}(s,\boldsymbol{a})]_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{S\prod_{i=1}^n A_i}$: a reward vector that represents the reward function for the i-th player at time step h.
- $\Pi_h^\pi \in \mathbb{R}^{S\times S\prod_{i=1}^n A_i}$: a projection matrix associated with time step h and a given joint policy $\pi = \{\pi_h\}_{h\in[H]}$ in the following form

$$\Pi_h^\pi = \begin{pmatrix} \pi_h(1)^\top & 0^\top & \cdots & 0^\top \\ 0^\top & \pi_h(2)^\top & \cdots & 0^\top \\ \vdots & \vdots & \ddots & \vdots \\ 0^\top & 0^\top & \cdots & \pi_h(S)^\top \end{pmatrix}, \tag{52}$$

  where we recall $\pi_h(s) = [\pi_h(s,\boldsymbol{a})]_{\boldsymbol{a}\in\mathcal{A}} \in \Delta(\mathcal{A})$ for all $s \in \mathcal{S}$ denote the joint policy vectors from all agents.
- $r_{i,h}^\pi \in \mathbb{R}^S$: a reward vector associated with the distribution of actions chosen by any joint policy $\pi = \{\pi_h\}_{h\in[H]}$ at time step h. Here, $r_{i,h}^\pi(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi_h(s)}[r_{i,h}(s,\boldsymbol{a})]$ for all $s \in \mathcal{S}$, or equivalently $r_{i,h}^\pi = \Pi_h^\pi r_{i,h}$ (see (52)).

- $P_h^0 \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$: the matrix of the nominal transition kernel at time step h, with $P_{h,s,\boldsymbol{a}}^0 \in \mathbb{R}^{1\times S}$ serves as the $(s,\boldsymbol{a})$-th row for any $(s,\boldsymbol{a}) \in \mathcal{S}\times\mathcal{A}$.
- $\widehat{P}_h^0 \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$: the matrix of the estimated nomimal transition kernel at time step h, with $\widehat{P}_{h,s,\boldsymbol{a}}^0 \in \mathbb{R}^{1\times S}$ serves as the $(s,\boldsymbol{a})$-th row for any $(s,\boldsymbol{a}) \in \mathcal{S}\times\mathcal{A}$.
- $P_{i,h}^V \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$, $\widehat{P}_{i,h}^V \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$: at time step h, those matrices represent the worst-case probability transition kernel within the i-th agent's uncertainty set around the nominal/estimated nominal transition kernel, associated with any vector $V \in \mathbb{R}^S$. As a result, we denote $P_{i,h,s,\boldsymbol{a}}^V$ (resp. $\widehat{P}_{i,h,s,\boldsymbol{a}}^V$) as the $(s,\boldsymbol{a})$-th row of the transition matrix $P_{i,h}^V$ (resp. $\widehat{P}_{i,h}^V$), defined by

$$P_{i,h,s,\boldsymbol{a}}^V = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}_\rho^{\sigma_i}(P_{h,s,\boldsymbol{a}}^0)}\mathcal{P}V, \qquad \text{and} \qquad \widehat{P}_{i,h,s,\boldsymbol{a}}^V = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}_\rho^{\sigma_i}(\widehat{P}_{h,s,\boldsymbol{a}}^0)}\mathcal{P}V. \tag{53a}$$

  Similarly, we define the corresponding probability transition matrices for some special value vectors that are useful: $P_{i,h}^{\pi,V} \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$, $P_{i,h}^{\pi,\widehat{V}} \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$, $\widehat{P}_{i,h}^{\pi,V} \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$ and $\widehat{P}_{i,h}^{\pi,\widehat{V}} \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$. Here, we already use the following short-hand notation:

$$\begin{aligned}
P_{i,h}^{\pi,V} &:= P_{i,h}^{V_{i,h+1}^{\pi,\sigma_i}} \quad \text{and} \quad P_{i,h,s,\boldsymbol{a}}^{\pi,V} := P_{i,h,s,\boldsymbol{a}}^{V_{i,h+1}^{\pi,\sigma_i}} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}_\rho^{\sigma_i}(P_{h,s,\boldsymbol{a}}^0)}\mathcal{P}V_{i,h+1}^{\pi,\sigma_i},\\
P_{i,h}^{\pi,\widehat{V}} &:= P_{i,h}^{\widehat{V}_{i,h+1}^{\pi,\sigma_i}} \quad \text{and} \quad P_{h,s,\boldsymbol{a}}^{\pi,\widehat{V}} := P_{h,s,\boldsymbol{a}}^{\widehat{V}_{i,h+1}^{\pi,\sigma_i}} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}_\rho^{\sigma_i}(P_{h,s,\boldsymbol{a}}^0)}\mathcal{P}\widehat{V}_{i,h+1}^{\pi,\sigma_i},\\
\widehat{P}_{i,h}^{\pi,V} &:= \widehat{P}_{i,h}^{V_{i,h+1}^{\pi,\sigma_i}} \quad \text{and} \quad \widehat{P}_{h,s,\boldsymbol{a}}^{\pi,V} := \widehat{P}_{h,s,\boldsymbol{a}}^{V_{i,h+1}^{\pi,\sigma_i}} = \operatorname{argmin}_{P\in\mathcal{U}_\rho^{\sigma_i}(\widehat{P}_{h,s,\boldsymbol{a}}^0)}PV_{i,h+1}^{\pi,\sigma_i},\\
\widehat{P}_{i,h}^{\pi,\widehat{V}} &:= \widehat{P}_{i,h}^{\widehat{V}_{i,h+1}^{\pi,\sigma_i}} \quad \text{and} \quad \widehat{P}_{h,s,\boldsymbol{a}}^{\pi,\widehat{V}} := \widehat{P}_{h,s,\boldsymbol{a}}^{\widehat{V}_{i,h+1}^{\pi,\sigma_i}} = \operatorname{argmin}_{P\in\mathcal{U}_\rho^{\sigma_i}(\widehat{P}_{h,s,\boldsymbol{a}}^0)}P\widehat{V}_{i,h+1}^{\pi,\sigma_i}.
\end{aligned} \tag{53b}$$

- $P_h^\pi \in \mathbb{R}^{S\times S}$, $\widehat{P}_h^\pi \in \mathbb{R}^{S\times S}$, $\underline{P}_{i,h}^{\pi,V} \in \mathbb{R}^{S\times S}$, $\underline{P}_{i,h}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$, $\underline{\widehat{P}}_{i,h}^{\pi,V} \in \mathbb{R}^{S\times S}$ and $\underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$: at time step h, those six *square* probability transition matrices w.r.t. a given joint policy π are defined by multiplying the projection matrix in (52) as below, resepctively:

$$\begin{aligned}
&\underline{P}_h^\pi := \Pi_h^\pi P_h^0, \qquad \underline{\widehat{P}}_h^\pi := \Pi_h^\pi \widehat{P}_h^0, \qquad \underline{P}_{i,h}^{\pi,V} := \Pi_h^\pi P_{i,h}^{\pi,V}, \qquad \underline{P}_{i,h}^{\pi,\widehat{V}} := \Pi_h^\pi P_{i,h}^{\pi,\widehat{V}},\\
&\underline{\widehat{P}}_{i,h}^{\pi,V} := \Pi_h^\pi \widehat{P}_{i,h}^{\pi,V}, \qquad \text{and} \qquad \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} := \Pi_h^\pi \widehat{P}_{i,h}^{\pi,\widehat{V}}.
\end{aligned} \tag{54}$$

We then introduce two notations of the variance. First, for any probability vector $P \in \mathbb{R}^{1\times S}$ and vector $V \in \mathbb{R}^S$, we denote the variance

$$\operatorname{Var}_P(V) := P(V\circ V) - (PV)\circ(PV). \tag{55}$$

Then in addition, for any transition kernel $P \in \mathbb{R}^{S\prod_{i=1}^n A_i \times S}$ and vector $V \in \mathbb{R}^S$, we denote $\mathsf{Var}_P(V) \in \mathbb{R}^{S\prod_{i=1}^n A_i}$ as a vector of variance whose $(s,\boldsymbol{a})$-th row of $\mathsf{Var}_P(V)$ is taken as

$$\mathsf{Var}_P(s,\boldsymbol{a}) := \operatorname{Var}_{P_{s,\boldsymbol{a}}}(V). \tag{56}$$

B.3. Preliminary facts of RMGs and empirical RMGs

Dual equivalence of robust Bellman operator with TV uncertainty set. Opportunely, when the prescribed uncertainty set is in a benign form (such as using TV distance as the divergence function), the robust Bellman operator can be computed efficiently by solving its dual formulation instead (Iyengar, 2005; Clavier et al., 2023; Shi et al., 2023). In particular, the following lemma describes the equivalence between the robust Bellman operator and its dual form due to strong duality in the case of TV distance.

Lemma B.1 (Lemma 4, Shi et al. (2023)). *Consider any TV uncertainty set $\mathcal{U}^\sigma(P) = \mathcal{U}_{\rho_{\mathsf{TV}}}^\sigma(P)$ associated with any probability vector $P \in \Delta(\mathcal{S})$, fixed uncertainty level $\sigma \in (0,1]$. For any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, recalling the definition of $[V]_\alpha$ in (24), one has*

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P)} \mathcal{P}V = \max_{\alpha\in[\min_s V(s), \max_s V(s)]} \left\{ P\left[V\right]_\alpha - \sigma\ \left(\alpha - \min_{s'}\left[V\right]_\alpha(s')\right)\right\}. \tag{57}$$

The above lemma ensures that the computation cost of applying robust Bellman operator is relatively the same as applying standard Bellman operator (Iyengar, 2005) up to some logarithmic factors.

Notations and facts of RMGs and empirical RMGs. First, recall that for any robust Markov game $\mathcal{MG}_{\mathsf{rob}} = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, \{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{1\le i\le n}, r, H\big\}$, according to robust Bellman equations in (13), one has for any joint policy $\pi : \mathcal{S}\times[H]\to\Delta(\mathcal{A})$ and any $(h,i,s,\boldsymbol{a})\in[H]\times[n]\times\mathcal{S}\times\mathcal{A}$:

$$Q_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a}) = r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})} PV_{i,h+1}^{\pi,\sigma_i}, \qquad \text{where } V_{i,h}^{\pi,\sigma_i}(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a})]. \tag{58}$$

Combined with the matrix notation in Appendix B.2, we arrive at

$$V_{i,h}^{\pi,\sigma_i} = r_{i,h}^{\pi} + \Pi_h^{\pi} \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P_h^0)} PV_{i,h+1}^{\pi,\sigma_i} = r_{i,h}^{\pi} + \underline{P}_{i,h}^{\pi,V} V_{i,h+1}^{\pi,\sigma_i}. \tag{59}$$

Then we denote the empirical robust Markov games based on the estimated nominal distribution $\widehat{P}^0$ constructed in (21) as $\widehat{\mathcal{MG}}_{\mathsf{rob}} = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, \{\mathcal{U}_\rho^{\sigma_i}(\widehat{P}^0)\}_{1\le i\le n}, r, H\big\}$. Analogous to (11), we can define the corresponding robust value function (resp. robust Q-function) of any joint policy π in $\widehat{\mathcal{MG}}_{\mathsf{rob}}$ as $\big\{\widehat{V}_{i,h}^{\pi,\sigma_i}\big\}_{1\le i\le n}$ (resp. $\big\{\widehat{Q}_{i,h}^{\pi,\sigma_i}\big\}_{1\le i\le n}$). In addition, similar to (12), we can define the maximum of the robust value function for each agent over $\widehat{\mathcal{MG}}_{\mathsf{rob}}$ as follows :

$$\forall s\in\mathcal{S}: \quad \widehat{V}_{i,h}^{\star,\pi_{-i},\sigma_i}(s) := \max_{\pi_i':\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} \widehat{V}_{i,h}^{\pi_i'\times\pi_{-i},\sigma_i}(s) = \max_{\pi_i':\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0)} \widehat{V}_{i,h}^{\pi_i'\times\pi_{-i},P}(s), \tag{60}$$

which can be achieved by at least one *robust best-response* policy for all $s\in\mathcal{S}$ simultaneously (Blanchet et al., 2024, Section A.2).

Moreover, applying the robust Bellman equation in (13) for the empirical RMG $\widehat{\mathcal{MG}}_{\mathsf{rob}}$, for any joint policy π,

$$\widehat{Q}_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a}) = r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P\widehat{V}_{i,h+1}^{\pi,\sigma_i}, \qquad \text{where } \widehat{V}_{i,h}^{\pi,\sigma_i}(s) = \mathbb{E}_{a\sim\pi_h(s)}[\widehat{Q}_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a})], \tag{61}$$

which combined with the matrix notations in Appendix B.2 leads to the matrix form of the robust Bellman equation:

$$\widehat{V}_{i,h}^{\pi,\sigma_i} = r_{i,h}^{\pi} + \Pi_h^{\pi} \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}_h^0)} P\widehat{V}_{i,h+1}^{\pi,\sigma_i} = r_{i,h}^{\pi} + \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i}. \tag{62}$$

Encouragingly, the above property of the robust Bellman equations ensure that the policy $\widehat{\pi}$ output by the proposed method DR-NVI (cf. Algorithm 1) is a robust-$\{$NE, CE, CCE$\}$ of the empirical RMG $\widehat{\mathcal{MG}}_{\mathsf{rob}}$ when executing different corresponding subroutines, summarized in the following lemma:

Lemma B.2. *The output policy $\widehat{\pi}$ by* DR-NVI *(cf. Algorithm 1) is a robust-$\{$NE, CE, CCE$\}$ of the empirical RMG $\widehat{\mathcal{MG}}_{\mathsf{rob}} = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, \{\mathcal{U}_\rho^{\sigma_i}(\widehat{P}^0)\}_{1\le i\le n}, r, H\big\}$ when executing different subroutine* Equilibrium $\in$ Compute$-\{$Nash, CE, CCE$\}$ *accordingly, namely*

$$\forall(i,h)\in[n]\times[H]: \quad \begin{cases} \widehat{V}_{i,h} = \widehat{V}_{i,h}^{\widehat{\pi},\sigma_i} = \widehat{V}_{i,h}^{\star,\widehat{\pi}_{-i},\sigma_i} & \textit{when } \mathsf{Equilibrium} = \mathsf{Compute-Nash} \\ \widehat{V}_{i,h} = \widehat{V}_{i,h}^{\widehat{\pi},\sigma_i} \ge \widehat{V}_{i,h}^{\star,\widehat{\pi}_{-i},\sigma_i} & \textit{when } \mathsf{Equilibrium} = \mathsf{Compute-CCE} \\ \widehat{V}_{i,h} = \widehat{V}_{i,h}^{\widehat{\pi},\sigma_i} \ge \max_{f_i\in\mathcal{F}_i} V_{i,h}^{f_i\diamond\widehat{\pi},\sigma_i} & \textit{when } \mathsf{Equilibrium} = \mathsf{Compute-CE}. \end{cases} \tag{63}$$

Proof. See Appendix C.3.1. □

C. Proof of Theorem 4.1

Before starting, let us introduce an essential lemma that characterize the difference between robust MGs and standard MGs. For each agent, the possible range of the robust value function shrinks as the uncertainty level σ_i of its own uncertainty set increases, shown below.

Lemma C.1. *Consider the uncertainty set $\mathcal{U}^{\sigma_i}(\cdot) = \mathcal{U}^{\sigma_i}_{\rho\text{-TV}}(\cdot)$ and any robust Markov game $\mathcal{MG}_{\mathsf{rob}} = \big\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, \{\mathcal{U}^{\sigma_i}(P)\}_{1\le i\le n}, r, H\big\}$. The robust value function $\{V^{\pi,\sigma_i}_{i,h}\}_{i\in[n],h\in[H]}$ associated with any joint policy π satisfies:*

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

Proof. See Appendix C.3.2 □

Equipped with the preceding lemma, we are now prepared to prove Theorem 4.1 for three different robust solution concepts, respectively.

C.1. Proof of learning robust NE/robust CCE

In this subsection, we focus on the two equilibrium concepts — robust NE and robust CCE. The proof is separated into several key steps as below.

Step 1: decomposing the error. Before proceeding, recall the goal is to prove that the output policy $\widehat{\pi}$ from Algorithm 1 is an ε-robust NE/CCE with corresponding subroutine (cf. line 4.1). Namely, $\widehat{\pi}\in\Delta(\mathcal{A}_1)\times\Delta(\mathcal{A}_2)\times\Delta(\mathcal{A}_n)$ is a product policy satisfies

$$\mathsf{gap}_{\mathsf{NE}}(\widehat{\pi}) \coloneqq \max_{s\in\mathcal{S},i\in[n]}\left\{V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1}(s) - V^{\widehat{\pi},\sigma_i}_{i,1}(s)\right\} \le \varepsilon \tag{64}$$

or $\widehat{\pi}\in\Delta(\mathcal{A})$ is a (possibly correlated) policy obeys

$$\mathsf{gap}_{\mathsf{CCE}}(\widehat{\pi}) \coloneqq \max_{s\in\mathcal{S}i\in[n]}\left\{V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1}(s) - V^{\widehat{\pi},\sigma_i}_{i,1}(s)\right\} \le \varepsilon. \tag{65}$$

We note that $\mathsf{gap}_{\mathsf{NE}}$ and $\mathsf{gap}_{\mathsf{CCE}}$ exhibit similar properties, differing only in the feasible set of policy $\widehat{\pi}$. So we consider them together.

To continue, we introduce the following best-response policy of the i-th player given other players policy $\widehat{\pi}_{-i}$:

$$\widetilde{\pi}^{\star}_i = \{\widetilde{\pi}^{\star}_{i,h}\}_{1\le h\le H} = \mathrm{argmax}_{\pi'_i\in\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} V^{\pi'_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1}, \tag{66}$$

which indicates that

$$V^{\widetilde{\pi}^{\star}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1} = V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1}. \tag{67}$$

Armed with above notations and facts, the term of interest $V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1} - V^{\widehat{\pi},\sigma_i}_{i,1}$ for any $i\in[n]$ can be decomposed as

$$\begin{aligned}
V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1} - V^{\widehat{\pi},\sigma_i}_{i,1} &= \left(V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1} - \widehat{V}^{\widetilde{\pi}^{\star}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1}\right) + \left(\widehat{V}^{\widetilde{\pi}^{\star}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1} - \widehat{V}^{\widehat{\pi},\sigma_i}_{i,1}\right) + \left(\widehat{V}^{\widehat{\pi},\sigma_i}_{i,1} - V^{\widehat{\pi},\sigma_i}_{i,1}\right)\\
&\overset{\text{(i)}}{\le} \left(V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1} - \widehat{V}^{\widetilde{\pi}^{\star}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1}\right) + \left(\widehat{V}^{\widetilde{\pi}^{\star}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1} - \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1}\right) + \left(\widehat{V}^{\widehat{\pi},\sigma_i}_{i,1} - V^{\widehat{\pi},\sigma_i}_{i,1}\right)\\
&\le \left(V^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1} - \widehat{V}^{\widetilde{\pi}^{\star}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1}\right) + \left(\widehat{V}^{\widehat{\pi},\sigma_i}_{i,1} - V^{\widehat{\pi},\sigma_i}_{i,1}\right)
\end{aligned} \tag{68}$$

where (i) holds by $\widehat{V}^{\widehat{\pi},\sigma_i}_{i,1} = \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1}$ (resp. $\widehat{V}^{\widehat{\pi},\sigma_i}_{i,1} \ge \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1}$) when the subroutine in line 4.1 is Compute − Nash (resp. Compute − CCE) implied by Lemma B.2, and the last inequality follows from $\widehat{V}^{\widetilde{\pi}^{\star}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1} \le \max_{\pi'_i\in\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} \widehat{V}^{\pi'_i\times\widehat{\pi}_{-i},\sigma_i}_{i,1} = \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,1}$ by definition.

Step 2: developing the recursion. We consider a more general form for any time step $h \in [H]$ and any joint policy π. Towards this, one has

$$\begin{aligned} V_{i,h}^{\pi,\sigma_i} - \widehat{V}_{i,h}^{\pi,\sigma_i} &\overset{\text{(i)}}{=} r_{i,h}^{\pi} + \Pi_h^{\pi} \inf_{P \in \mathcal{U}^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})} P V_{i,h+1}^{\pi,\sigma_i} - \left(r_{i,h}^{\pi} + \Pi_h^{\pi} \inf_{P \in \mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right) \\ &\overset{\text{(ii)}}{=} \underline{P}_{i,h}^{\pi,V} V_{i,h+1}^{\pi,\sigma_i} - \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} \qquad (69) \\ &= \left(\underline{P}_{i,h}^{\pi,V} V_{i,h+1}^{\pi,\sigma_i} - \underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right) + \left(\underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} - \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right) \\ &\overset{\text{(iii)}}{\leq} \underline{P}_{i,h}^{\pi,\widehat{V}} \left(V_{i,h+1}^{\pi,\sigma_i} - \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right) + \underbrace{\left| \underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} - \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right|}_{=: a_{i,h}^{\pi}} \qquad (70) \end{aligned}$$

where (i) and (ii) hold by the matrix version of robust Bellman consistency equations in (59) and (62), and (iii) follows from the observation

$$\underline{P}_{i,h}^{\pi,V} V_{i,h+1}^{\pi,\sigma_i} \leq \underline{P}_{i,h}^{\pi,\widehat{V}} V_{i,h+1}^{\pi,\sigma_i}$$

due to the definition of $\underline{P}_{i,h}^{\pi,V} = \Pi_h^{\pi} \arg\min_{P \in \mathcal{U}^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})} P V_{i,h+1}^{\pi,\sigma_i} \leq \Pi_h^{\pi} \arg\min_{P \in \mathcal{U}^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})} P \widehat{V}_{i,h+1}^{\pi,\sigma_i}$ (cf. (53) and (54)).

Recursively applying (70) leads to

$$\begin{aligned} & V_{i,h}^{\pi,\sigma_i} - \widehat{V}_{i,h}^{\pi,\sigma_i} \\ & \leq \underline{P}_{i,h}^{\pi,\widehat{V}} \underline{P}_{i,h+1}^{\pi,\widehat{V}} \left(V_{i,h+2}^{\pi,\sigma_i} - \widehat{V}_{i,h+2}^{\pi,\sigma_i} \right) + \underline{P}_{i,h}^{\pi,\widehat{V}} \left| \underline{P}_{i,h+1}^{\pi,\widehat{V}} \widehat{V}_{i,h+2}^{\pi,\sigma_i} - \underline{\widehat{P}}_{i,h+1}^{\pi,\widehat{V}} \widehat{V}_{i,h+2}^{\pi,\sigma_i} \right| + \left| \underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} - \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right| \\ & \leq \cdots \leq \sum_{j=h}^{H} \left(\prod_{k=h}^{j-1} \underline{P}_{i,k}^{\pi,\widehat{V}} \right) a_{i,j}^{\pi}, \qquad (71) \end{aligned}$$

where the last inequality holds by adopting the following notations

$$\left(\prod_{k=h}^{h-1} \underline{P}_{i,k}^{\pi,\widehat{V}} \right) = I \quad \text{and} \quad \left(\prod_{k=h}^{j-1} \underline{P}_{i,k}^{\pi,\widehat{V}} \right) = \underline{P}_{i,h}^{\pi,\widehat{V}} \cdot \underline{P}_{i,h+1}^{\pi,\widehat{V}} \cdots \underline{P}_{i,j-1}^{\pi,\widehat{V}}. \qquad (72)$$

Next, similar to (70), we can achieve

$$\begin{aligned} \widehat{V}_{i,h}^{\pi,\sigma_i} - V_{i,h}^{\pi,\sigma_i} &\overset{\text{(i)}}{=} \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} - \underline{P}_{i,h}^{\pi,V} V_{i,h+1}^{\pi,\sigma_i} \\ &= \left(\underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} - \underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right) + \left(\underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} - \underline{P}_{i,h}^{\pi,V} V_{i,h+1}^{\pi,\sigma_i} \right) \\ &\leq \underline{P}_{i,h}^{\pi,V} \left(\widehat{V}_{i,h+1}^{\pi,\sigma_i} - V_{i,h+1}^{\pi,\sigma_i} \right) + \left| \underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} - \underline{\widehat{P}}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi,\sigma_i} \right| \qquad (73) \end{aligned}$$

where (i) holds by (69), and the last inequality follows from the fact $\underline{P}_{i,h}^{\pi,\widehat{V}} \widehat{V}_{i,h+1}^{\pi} \leq \underline{P}_{i,h}^{\pi,V} \widehat{V}_{i,h+1}^{\pi}$ (see the definition of $\underline{P}_{i,h}^{\pi,\widehat{V}}$, i.e., (53) and (54)).

Then following the routine of achieving (71), we arrive at

$$\widehat{V}_{i,h}^{\pi,\sigma_i} - V_{i,h}^{\pi,\sigma_i} \leq \sum_{j=h}^{H} \left(\prod_{k=h}^{j-1} \underline{P}_{i,k}^{\pi,V} \right) a_{i,j}^{\pi}. \qquad (74)$$

Summing up (71) and (74), one has for any joint policy π,

$$\left| \widehat{V}_{i,h}^{\pi,\sigma_i} - V_{i,h}^{\pi,\sigma_i} \right| \leq \max\{ V_{i,h}^{\pi,\sigma_i} - \widehat{V}_{i,h}^{\pi,\sigma_i}, \widehat{V}_{i,h}^{\pi,\sigma_i} - V_{i,h}^{\pi,\sigma_i} \}$$

$$\leq \max\left\{\sum_{j=h}^{H}\left(\prod_{k=h}^{j-1}\underline{P}_{i,k}^{\pi,\widehat{V}}\right)a_{i,j}^{\pi},\sum_{j=h}^{H}\left(\prod_{k=h}^{j-1}\underline{P}_{i,k}^{\pi,V}\right)a_{i,j}^{\pi}\right\}, \tag{75}$$

where the $\max$ operator is taken entry-wise for the vectors.

To continue, we introduce an important concentration result about the value estimation error as follows:

Lemma C.2. *Consider any $\delta\in(0,1)$. With probability at least $1-\delta$, one has for any joint policy π,*

$$\begin{aligned}\forall(h,i)\in[H]\times[n]:\quad a_{i,h}^{\pi}&=\left|\underline{P}_{i,h}^{\pi,\widehat{V}}\widehat{V}_{i,h+1}^{\pi,\sigma_i}-\widehat{\underline{P}}_{i,h}^{\pi,\widehat{V}}\widehat{V}_{i,h+1}^{\pi,\sigma_i}\right|\\&\leq 2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\sqrt{\mathsf{Var}_{\underline{P}_h^{\pi}}(\widehat{V}_{i,h+1}^{\pi})}+\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})H}{N}1\\&\leq 3\sqrt{\frac{H^2\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}1\end{aligned} \tag{76}$$

where $\mathsf{Var}_{\underline{P}_h^{\pi}}(\cdot)$ is defined in (56).

Proof. See Appendix C.3.3. □

Step 3: controlling the first term in (75). Let us introduce some additional notations for convenience. Recall e_s denote a S-dimensional standard basis supported on the s-th element. We denote

$$d_h^h=e_s\quad\text{and}\quad d_h^j=e_s^{\top}\left(\prod_{k=h}^{j-1}\underline{P}_{i,k}^{\pi,\widehat{V}}\right)\quad\forall j=h+1,\cdots,H. \tag{77}$$

Armed with above notations and facts, for any $s\in\mathcal{S}$, we have

$$\begin{aligned}V_{i,h}^{\pi,\sigma_i}(s)-\widehat{V}_{i,h}^{\pi,\sigma_i}(s)&=\left\langle e_s,V_{i,h}^{\pi,\sigma_i}-\widehat{V}_{i,h}^{\pi,\sigma_i}\right\rangle=\sum_{j=h}^{H}\left\langle d_h^j,a_{i,j}^{\pi}\right\rangle\\&\leq\sum_{j=h}^{H}\left\langle d_h^j,\left(2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\sqrt{\mathsf{Var}_{\underline{P}_j^{\pi}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})}+\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})H}{N}1\right)\right\rangle\\&\leq\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})H^2}{N}+2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\sum_{j=h}^{H}\left\langle d_h^j,\sqrt{\mathsf{Var}_{\underline{P}_j^{\pi}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})}\right\rangle\\&\overset{\text{(i)}}{\leq}\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})H^2}{N}+2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\sqrt{H\sum_{j=h}^{H}\left\langle d_h^j,\mathsf{Var}_{\underline{P}_j^{\pi}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right\rangle}\\&\leq\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})H^2}{N}+\underbrace{2\sqrt{\frac{H\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\sqrt{\sum_{j=h}^{H}\left\langle d_h^j,\mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right\rangle}}_{=:\mathcal{B}_1}\\&\quad+\underbrace{2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\sqrt{H\sum_{j=h}^{H}\left\langle d_h^j,\left|\mathsf{Var}_{\underline{P}_j^{\pi}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})-\mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right|\right\rangle}}_{=:\mathcal{B}_2}\end{aligned} \tag{78}$$

where (i) holds by the Cauchy-Schwarz inequality.

Then we control the two main terms in (78) separately.

- **Controlling $\mathcal{B}_1$.** To begin with, we introduce the following lemma about $\sum_{j=h}^{H}\left\langle d_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right\rangle$ whose proof is postponed to Appendix C.3.4.

Lemma C.3. *Consider any $\delta\in(0,1)$. With probability at least $1-\delta$, one has for any joint policy π,*

$$\forall(h,i)\in[H]\times[n]:\quad \sum_{j=h}^{H}\left\langle d_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right\rangle \le 3H\left(\max_{s\in\mathcal{S}}\widehat{V}_{i,j+1}^{\pi,\sigma_i}(s)-\min_{s\in\mathcal{S}}\widehat{V}_{i,j+1}^{\pi,\sigma_i}(s)\right)\left(1+2H\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\right). \tag{79}$$

Applying Lemma C.3 to $\mathcal{B}_1$ in (78), we arrive at

$$\begin{aligned}
\mathcal{B}_1 &= 2\sqrt{\frac{H\log\left(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta}\right)}{N}}\sqrt{\sum_{j=h}^{H}\left\langle d_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right\rangle}\\
&\le 2\sqrt{\frac{H\log\left(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta}\right)}{N}}\sqrt{3H\left(\max_{s\in\mathcal{S}}\widehat{V}_{i,j+1}^{\pi,\sigma_i}(s)-\min_{s\in\mathcal{S}}\widehat{V}_{i,j+1}^{\pi,\sigma_i}(s)\right)\left(1+2H\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta}\right)}{N}}\right)}\\
&\overset{\text{(i)}}{\le} 2\sqrt{\frac{3H^2\log\left(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta}\right)}{N}\min\left\{\frac{1}{\sigma_i},H-h+1\right\}\left(1+2H\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta}\right)}{N}}\right)}\\
&\le 6\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log\left(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta}\right)}{N}},
\end{aligned} \tag{80}$$

where (i) holds by applying Lemma C.3.2, and the last inequality follows by taking $N\ge 4H^2\log\left(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta}\right)$.

- **Controlling $\mathcal{B}_2$.** We introduce another lemma; refer to the proof in Appendix C.3.5.

Lemma C.4. *Consider the standard RMG $\mathcal{MG}=\left\{\mathcal{S},\{\mathcal{A}_i\}_{1\le i\le n},\{\mathcal{U}^{\sigma_i}(P^0)\}_{1\le i\le n},r,H\right\}$ and empirical RMG $\mathcal{MG}_{\mathsf{rob}}=\left\{\mathcal{S},\{\mathcal{A}_i\}_{1\le i\le n},\{\mathcal{U}^{\sigma_i}(\widehat{P}^0)\}_{1\le i\le n},r,H\right\}$. Considering any joint policy π, any transition kernel $P'\in\mathbb{R}^S$ and any $\widetilde{P}\in\mathbb{R}^S$ obeying $\widetilde{P}\in\mathcal{U}^{\sigma_i}(P)$, one has*

$$\forall(i,j)\in[n]\times[H]:\quad \left|\mathsf{Var}_{P'}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})-\mathsf{Var}_{\widetilde{P}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right|\le\min\left\{\frac{1}{\sigma_i},H-h+1\right\}, \tag{81a}$$

$$\left|\mathsf{Var}_{P'}(V_{i,j+1}^{\pi,\sigma_i})-\mathsf{Var}_{\widetilde{P}}(V_{i,j+1}^{\pi,\sigma_i})\right|\le\min\left\{\frac{1}{\sigma_i},H-h+1\right\}. \tag{81b}$$

Armed with above lemma, we observe that

$$\begin{aligned}
\left|\mathsf{Var}_{\underline{P}_j^{\pi}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})-\mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right| &\overset{\text{(i)}}{=}\left|\Pi_j^{\pi}\left(\mathsf{Var}_{P_j^0}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})-\mathsf{Var}_{P_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right)\right|\\
&\overset{\text{(ii)}}{\le}\left\|\mathsf{Var}_{P_j^0}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})-\mathsf{Var}_{P_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right\|_{\infty}1\\
&\le\min\left\{\frac{1}{\sigma_i},H-h+1\right\}1,
\end{aligned} \tag{82}$$

where (i) and (ii) follows from the matrix notations Π_j^{π} (cf (52)) and $\underline{P}_j^{\pi},\underline{P}_{i,j}^{\pi,\widehat{V}}$ (cf (54)), and the last inequality holds by applying Lemma C.4 with $P'=P_{j,s,\boldsymbol{a}}^0,\widetilde{P}=P_{i,j,s,\boldsymbol{a}}^{\pi,\widehat{V}}$ for all $(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}$.

Plugging back (82) to (78), it can be verified that

$$\begin{aligned}\mathcal{B}_2 &= 2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\sqrt{H\sum_{j=h}^{H}\left\langle d_h^j, \left|\mathsf{Var}_{\underline{P}_j^\pi}(\widehat{V}_{i,j+1}^{\pi,\sigma_i}) - \mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})\right|\right\rangle}\\ &\le 2\sqrt{\frac{H\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\sqrt{\sum_{j=h}^{H}\left\langle d_h^j, \min\left\{\frac{1}{\sigma_i}, H-h+1\right\}1\right\rangle}\\ &\le 2\sqrt{\frac{H^2\min\{1/\sigma_i, H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}. \end{aligned} \tag{83}$$

Consequently, combining (80) and (83), (78) can be bounded by

$$\begin{aligned}V_{i,h}^{\pi,\sigma_i}(s) - \widehat{V}_{i,h}^{\pi,\sigma_i}(s) &\le \frac{\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})H^2}{N} + 6\sqrt{\frac{H^2\min\{1/\sigma_i, H\}\log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)}{N}}\\ &\quad + 2\sqrt{\frac{H^2\min\{1/\sigma_i, H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\\ &\le 9\sqrt{\frac{H^2\min\{1/\sigma_i, H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}, \end{aligned} \tag{84}$$

where the last inequality holds by taking $N \ge 4H^2\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})$.

Step 4: controlling the second term in (75). To do so, similar to (77), we define

$$w_h^h = e_s \quad \text{and} \quad w_h^j = e_s^\top\left(\prod_{k=h}^{j-1}\underline{P}_{i,k}^{\pi,V}\right) \quad \forall j = h+1,\cdots,H. \tag{85}$$

With the above notations in mind, following the routine of (78) gives: for any $s\in\mathcal{S}$,

$$\begin{aligned}&\widehat{V}_{i,h}^{\pi,\sigma_i}(s) - V_{i,h}^{\pi,\sigma_i}(s)\\ &\le \frac{\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})H^2}{N} + 2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\sum_{j=h}^{H}\left\langle w_h^j, \sqrt{\mathsf{Var}_{\underline{P}_j^\pi}(\widehat{V}_{i,j+1}^{\pi,\sigma_i})}\right\rangle\\ &\overset{\text{(i)}}{\le} \frac{\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})H^2}{N} + 2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\sum_{j=h}^{H}\Big\langle w_h^j,\\ &\quad\left(\sqrt{\left|\mathsf{Var}_{\underline{P}_j^\pi}(\widehat{V}_{i,j+1}^{\pi,\sigma_i} - V_{i,j+1}^{\pi,\sigma_i})\right|} + \sqrt{\left|\mathsf{Var}_{\underline{P}_j^\pi}(V_{i,j+1}^{\pi,\sigma_i}) - \mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}(V_{i,j+1}^{\pi,\sigma_i})\right|} + \sqrt{\mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}(V_{i,j+1}^{\pi,\sigma_i})}\right)\Big\rangle\\ &\le \frac{H^2\log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)}{N} + \underbrace{2\sqrt{\frac{H\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\sqrt{\sum_{j=h}^{H}\left\langle w_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}(V_{i,j+1}^{\pi,\sigma_i})\right\rangle}}_{=:\mathcal{B}_3}\\ &\quad\underbrace{+ 2\sqrt{\frac{H\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\sqrt{\sum_{j=h}^{H}\left\langle w_h^j, \left|\mathsf{Var}_{\underline{P}_j^\pi}(V_{i,j+1}^{\pi,\sigma_i}) - \mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}(V_{i,j+1}^{\pi,\sigma_i})\right|\right\rangle}}_{=:\mathcal{B}_4}\\ &\quad\underbrace{+ 2\sqrt{\frac{H\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\sqrt{\sum_{j=h}^{H}\left\langle w_h^j, \left|\mathsf{Var}_{\underline{P}_j^\pi}(\widehat{V}_{i,j+1}^{\pi,\sigma_i} - V_{i,j+1}^{\pi,\sigma_i})\right|\right\rangle}}_{=:\mathcal{B}_5}, \end{aligned} \tag{86}$$

where (i) holds by the triangle inequality and the elementary inequality $\sqrt{\mathsf{Var}_P(V+V')} \le \sqrt{\mathsf{Var}_P(V)} + \sqrt{\mathsf{Var}_P(V')}$ for any transition kernel $P \in \mathbb{R}^S$ and vectors $V, V' \in \mathbb{R}^S$, and the last inequality follows from applying the Cauchy-Schwarz inequality to those terms.

We can control the three main terms in (86) separately as below:

- **Controlling $\mathcal{B}_3$.** First, we introduce the following lemma for $\sum_{j=h}^{H} \left\langle w_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}(V_{i,j+1}^{\pi,\sigma_i}) \right\rangle$.

  Lemma C.5. *Consider any $\delta \in (0,1)$. For any joint policy π, with probability at least $1-\delta$,*

  $$\forall (h,i) \in [H] \times [n]: \quad \sum_{j=h}^{H} \left\langle w_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,\widehat{V}}}(V_{i,j+1}^{\pi,\sigma_i}) \right\rangle \le 3H \left(\max_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) - \min_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) \right). \tag{87}$$

  Proof. See Appendix C.3.6. □

  Then applying Lemma C.5 yields

  $$\begin{aligned}
  \mathcal{B}_3 &= 2\sqrt{\frac{H \log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)}{N}} \sqrt{\sum_{j=h}^{H} \left\langle w_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}(V_{i,j+1}^{\pi,\sigma_i}) \right\rangle} \\
  &\le 2\sqrt{\frac{H \log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)}{N}} \sqrt{3H \left(\max_{s\in\mathcal{S}} \widehat{V}_{i,h}^{\pi,\sigma_i}(s) - \min_{s\in\mathcal{S}} \widehat{V}_{i,h}^{\pi,\sigma_i}(s) \right)} \\
  &\le 4\sqrt{\frac{H^2 \min\{1/\sigma_i, H\} \log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)}{N}},
  \end{aligned} \tag{88}$$

  where the last inequality follows from Lemma C.1.

- **Controlling $\mathcal{B}_4$ and $\mathcal{B}_5$.** First, it is easily verified that $\mathcal{B}_4$ can be controlled as the same as that for $\mathcal{B}_2$ (see (83)) by applying Lemma (81b), namely

  $$\mathcal{B}_4 \le 2\sqrt{\frac{H^2 \min\{1/\sigma_i, H\} \log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}. \tag{89}$$

  Then the remainder of the proof shall focus on $\mathcal{B}_5$. Recalling the definition in (86), one has

  $$\begin{aligned}
  \mathcal{B}_5 &= 2\sqrt{\frac{H \log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}} \sqrt{\sum_{j=h}^{H} \left\langle w_h^j, \left| \mathsf{Var}_{\underline{P}_j^{\pi}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i} - V_{i,j+1}^{\pi,\sigma_i}) \right| \right\rangle} \\
  &\le 2\sqrt{\frac{H^2 \log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}} \sqrt{\max_{h\le j\le H} \left\| \mathsf{Var}_{\underline{P}_j^{\pi}}(\widehat{V}_{i,j+1}^{\pi,\sigma_i} - V_{i,j+1}^{\pi,\sigma_i}) \right\|_\infty} \\
  &\le 2\sqrt{\frac{H^2 \log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}} \max_{h\le j\le H} \left\| \widehat{V}_{i,j+1}^{\pi,\sigma_i} - V_{i,j+1}^{\pi,\sigma_i} \right\|_\infty.
  \end{aligned} \tag{90}$$

Summing up (88), (89), and (90) and inserting back to (86), we conclude

$$\begin{aligned}
&\widehat{V}_{i,h}^{\pi,\sigma_i}(s) - V_{i,h}^{\pi,\sigma_i}(s) \\
&\le \frac{\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}) H^2}{N} + 4\sqrt{\frac{H^2 \min\{1/\sigma_i, H\} \log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}
\end{aligned}$$

$$
\begin{aligned}
&+2\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}+2\sqrt{\frac{H^2\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\max_{h\le j\le H}\left\|\widehat{V}_{i,j+1}^{\pi,\sigma_i}-V_{i,j+1}^{\pi,\sigma_i}\right\|_\infty\\
&\le 7\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}1\\
&\quad+2\sqrt{\frac{H^2\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\max_{h\le j\le H}\left\|\widehat{V}_{i,j+1}^{\pi,\sigma_i}-V_{i,j+1}^{\pi,\sigma_i}\right\|_\infty 1, \qquad (91)
\end{aligned}
$$

as long as $N\ge H^2\log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)$.

Step 5: summing up the results. Inserting (84) and (91) back into (75), we observe that

$$
\begin{aligned}
&\left|\widehat{V}_{i,h}^{\pi,\sigma_i}-V_{i,h}^{\pi,\sigma_i}\right|\\
&\le\max\left\{V_{i,h}^{\pi,\sigma_i}-\widehat{V}_{i,h}^{\pi,\sigma_i},\widehat{V}_{i,h}^{\pi,\sigma_i}-V_{i,h}^{\pi,\sigma_i}\right\}\\
&\le\max\Bigg\{9\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}1,\\
&\quad 7\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}1+2\sqrt{\frac{H^2\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\max_{h\le j\le H}\left\|\widehat{V}_{i,j+1}^{\pi,\sigma_i}-V_{i,j+1}^{\pi,\sigma_i}\right\|_\infty 1\Bigg\}, \qquad (92)
\end{aligned}
$$

which indicates

$$
\begin{aligned}
&\max_{h\in[H]}\left\|\widehat{V}_{i,h}^{\pi,\sigma_i}-V_{i,h}^{\pi,\sigma_i}\right\|_\infty\\
&\le 9\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}1+2\sqrt{\frac{H^2\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}\max_{h\in[H]}\left\|\widehat{V}_{i,h+1}^{\pi,\sigma_i}-V_{i,h+1}^{\pi,\sigma_i}\right\|_\infty\\
&\overset{\text{(i)}}{\le} 9\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}1+\frac{1}{2}\max_{h\in[H]}\left\|\widehat{V}_{i,h}^{\pi,\sigma_i}-V_{i,h}^{\pi,\sigma_i}\right\|_\infty\\
&\le 18\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}, \qquad (93)
\end{aligned}
$$

where (i) holds by taking $N\ge 16H^2\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})$ and invoking the basic fact that $\widehat{V}_{i,H+1}^{\pi,\sigma_i}=V_{i,H+1}^{\pi,\sigma_i}=0$.

Finally, we complete the proof by showing that the performance gap in (68) is bounded by

$$
\begin{aligned}
V_{i,1}^{\star,\widehat{\pi}_{-i}}-V_{i,1}^{\widehat{\pi}}&\le\left(V_{i,1}^{\star,\widehat{\pi}_{-i}}-\widehat{V}_{i,1}^{\widetilde{\pi}_i^\star\times\widehat{\pi}_{-i}}\right)+\left(\widehat{V}_{i,1}^{\widehat{\pi}}-V_{i,1}^{\widehat{\pi}}\right)\\
&\le\left\|V_{i,1}^{\star,\widehat{\pi}_{-i}}-\widehat{V}_{i,1}^{\widetilde{\pi}_i^\star\times\widehat{\pi}_{-i}}\right\|_\infty 1+\left\|\widehat{V}_{i,1}^{\widehat{\pi}}-V_{i,1}^{\widehat{\pi}}\right\|_\infty 1\\
&\le\max_{h\in[H]}\left\|V_{i,h}^{\star,\widehat{\pi}_{-i}}-\widehat{V}_{i,h}^{\widetilde{\pi}_i^\star\times\widehat{\pi}_{-i}}\right\|_\infty 1+\max_{h\in[H]}\left\|\widehat{V}_{i,h}^{\widehat{\pi}}-V_{i,h}^{\widehat{\pi}}\right\|_\infty 1\\
&\le 36\sqrt{\frac{H^2\min\{1/\sigma_i,H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}1, \qquad (94)
\end{aligned}
$$

where the last inequality holds by applying (93) to two different cases when $\pi=\widetilde{\pi}_i^\star\times\widehat{\pi}_{-i}$ or $\pi=\widehat{\pi}$, respectively.

As a result, to achieve $\max_{s\in\mathcal{S},i\in[n]}\left\{V_{i,1}^{\star,\widehat{\pi}_{-i},\sigma_i}(s)-V_{i,1}^{\widehat{\pi},\sigma_i}(s)\right\}\le\varepsilon$ with probability at least $1-\delta$, we require the total number of samples

$$
N_{\mathsf{all}}=HS\prod_{i\in[n]}A_iN\ge\frac{C_1SH^3\prod_{1\le i\le n}A_i\log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)}{\varepsilon^2}\min\left\{H,\frac{1}{\min_{1\le i\le n}\sigma_i}\right\}
$$

$$\geq \frac{C_0 S H^3 \prod_{1\leq i\leq n} A_i \log\left(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}\right)}{\varepsilon^2} \min\left\{H, \frac{1}{\min_{1\leq i\leq n}\sigma_i}\right\}$$
$$+ 16H^3 S \prod_{i\in[n]} A_i \log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta}), \tag{95}$$

providing $C_1 > C_0$ are larger enough universal constant, and $\varepsilon \leq \sqrt{\min\left\{H, \frac{1}{\min_{1\leq i\leq n}\sigma_i}\right\}}$.

C.2. Proof of learning robust CE

This section is analogous to the proof for learning robust NE/CCE in Appendix C.1.

The goal is to prove that the policy $\widehat{\pi}$ output from Algorithm 1 is an ε-robust CE when executing subroutine $\mathsf{Compute-CE}(\cdot)$ for line 4.1, i.e.,

$$\mathsf{gap}_{\mathsf{CE}}(\widehat{\pi}) = \max_{s\in\mathcal{S}, 1\leq i\leq n}\left\{\max_{f_i\in\mathcal{F}_i} V_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i}(s) - V_{i,1}^{\widehat{\pi},\sigma_i}(s)\right\} \leq \varepsilon. \tag{96}$$

So we define the following best perturbation policy of the i-th player as

$$\overline{\pi}_i^\star = \{\overline{\pi}_{i,h}^\star\}_{1\leq h\leq H} = \left(\mathrm{argmax}_{f_i\in\mathcal{F}_i} V_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i}\right)\diamond\widehat{\pi} \tag{97}$$

which leads to

$$V_{i,1}^{\overline{\pi}_i^\star,\sigma_i} = \max_{f_i\in\mathcal{F}_i} V_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i}. \tag{98}$$

With above notations in mind, for any $1\leq i\leq n$, the term of interest can be decomposed as

$$\begin{aligned}
\max_{f_i\in\mathcal{F}_i} V_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i} - V_{i,1}^{\widehat{\pi},\sigma_i} &= \left(V_{i,1}^{\overline{\pi}_i^\star,\sigma_i} - \widehat{V}_{i,1}^{\overline{\pi}_i^\star,\sigma_i}\right) + \left(\widehat{V}_{i,1}^{\overline{\pi}_i^\star,\sigma_i} - \widehat{V}_{i,1}^{\widehat{\pi},\sigma_i}\right) + \left(\widehat{V}_{i,1}^{\widehat{\pi},\sigma_i} - V_{i,1}^{\widehat{\pi},\sigma_i}\right) \\
&\overset{\text{(i)}}{\leq} \left(V_{i,1}^{\overline{\pi}_i^\star,\sigma_i} - \widehat{V}_{i,1}^{\overline{\pi}_i^\star,\sigma_i}\right) + \left(\widehat{V}_{i,1}^{\overline{\pi}_i^\star,\sigma_i} - \max_{f_i\in\mathcal{F}_i}\widehat{V}_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i}\right) + \left(\widehat{V}_{i,1}^{\widehat{\pi},\sigma_i} - V_{i,1}^{\widehat{\pi},\sigma_i}\right) \\
&\leq \left(V_{i,1}^{\overline{\pi}_i^\star,\sigma_i} - \widehat{V}_{i,1}^{\overline{\pi}_i^\star,\sigma_i}\right) + \left(\widehat{V}_{i,1}^{\widehat{\pi},\sigma_i} - V_{i,1}^{\widehat{\pi},\sigma_i}\right)
\end{aligned} \tag{99}$$

where (i) holds by $\widehat{V}_{i,1}^{\widehat{\pi},\sigma_i} \geq \max_{f_i\in\mathcal{F}_i} V_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i}$ when the subroutine in line 4.1 is $\mathsf{Compute-CE}(\cdot)$ implied by Lemma B.2, and the last inequality follows from $\widehat{V}_{i,1}^{\overline{\pi}_i^\star,\sigma_i} = \widehat{V}_{i,1}^{\overline{f}_i\diamond\widehat{\pi},\sigma_i} \leq \max_{f_i\in\mathcal{F}_i}\widehat{V}_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i}$ for some $\overline{f}_i\in\mathcal{F}_i$.

Observing that (99) is similar to (68), it can be verified that following the same pipeline routine and the same facts developed from Step 2 to Step 5 in Appendix C.1, we can achieve similar results as below:

$$\forall i\in[n]: \quad \max_{f_i\in\mathcal{F}_i} V_{i,1}^{f_i\diamond\widehat{\pi},\sigma_i} - V_{i,1}^{\widehat{\pi},\sigma_i} \leq 36\sqrt{\frac{H^2\min\{1/\sigma_i, H\}\log(\frac{18S\prod_{i=1}^n A_i nHN}{\delta})}{N}}1, \tag{100}$$

which yields (95) and complete the proof. We omit the details here for conciseness.

C.3. Proof of the auxiliary lemmas

C.3.1. Proof of Lemma B.2

We will prove each line of (63) separately with an induction argument. Note that Blanchet et al. (2023) provides the proof of the first line of (63) for robust NE. For completeness, we offer the whole proof for all of the three robust solution concepts (including robust-NE).

Proof for robust NE. First, we focus on the first line of (63) and provide the following induction argument:

- *Base case when $h = H$.* Note that $\widehat{V}_{i,H+1}^{\pi,\sigma_i} = 0$ for all $i \in [n]$ are satisfied by definition. As a result, the robust Q-function for any joint policy π and the estimate from Algorithm 1 satisfy

$$\forall (i,s,a) \in [n] \times \mathcal{S} \times \mathcal{A}: \quad \widehat{Q}_{i,H}^{\pi,\sigma_i}(s,\boldsymbol{a}) = r_{i,H}(s,\boldsymbol{a}) \quad \text{and} \quad \widehat{Q}_{i,H}(s,\boldsymbol{a}) = r_{i,H}(s,\boldsymbol{a}) \tag{101}$$

which directly leads to

$$\widehat{V}_{i,H}^{\pi,\sigma_i} = \widehat{V}_{i,H} \tag{102}$$

and the output π_H obeying

$$\forall s \in \mathcal{S}: \quad \widehat{\pi}_H(\cdot \,|\, s) \leftarrow \mathsf{Compute-Nash}\left(r_{1,H}(s,\boldsymbol{a}), r_{2,H}(s,\boldsymbol{a}), \cdots, r_{n,H}(s,\boldsymbol{a})\right). \tag{103}$$

Consequently, invoking line 4.1 of Algorithm 1 gives that for all $s \in \mathcal{S}$,

$$\widehat{V}_{i,H}(s) = \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_H(s)}\left[\widehat{Q}_{i,H}(s,\boldsymbol{a})\right] \overset{\text{(i)}}{=} \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_H(s)}\left[\widehat{Q}_{i,H}^{\widehat{\pi},\sigma_i}(s,\boldsymbol{a})\right] \overset{\text{(ii)}}{=} \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_H(s)}[r_{i,H}(s,\boldsymbol{a})] \tag{104}$$

$$\overset{\text{(iii)}}{=} \max_{\widetilde{\pi}_{i,H}(s)\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,H}(s)\times\widehat{\pi}_{-i,H}(s)}[r_{i,H}(s,\boldsymbol{a})] \tag{105}$$

$$\overset{\text{(iv)}}{=} \max_{\widetilde{\pi}_{i,H}(s)\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,H}(s)\times\widehat{\pi}_{-i,H}(s)}\left[\widehat{Q}_{i,H}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}(s,\boldsymbol{a})\right]$$

$$= \max_{\widetilde{\pi}_i:\mathcal{S}\times[H]\to\in\Delta(\mathcal{A}_i)} \widehat{V}_{i,H}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}(s) = \widehat{V}_{i,H}^{\star,\widehat{\pi}_{-i},\sigma_i}, \tag{106}$$

where (i) and (ii) hold by (101), (iii) arises from the definition of robust-NE (see (103)) associated with $\{r_{i,H}\}_{i\in[n]}$, (iv) holds by applying (101) for policy $\pi = \widetilde{\pi}_i \times \widehat{\pi}_{-i}$, and the penultimate equality follows from the fact that only the policy of the time step H will influence $\widehat{V}_{i,H}^{\pi,\sigma_i}(s)$ due to Markov property. Thus we complete the proof for the base case.

- *Induction.* To continue, suppose the first line in (63) holds for step $h+1$, we shall proof that it also holds for time step h. To proceed, applying the robust Bellman equation in (61) for the TV uncertainty set $\mathcal{U}^{\sigma_i}(\cdot)$, we observe that

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{Q}_{i,h}^{\widehat{\pi},\sigma_i}(s,\boldsymbol{a}) = r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P\widehat{V}_{i,h+1}^{\widehat{\pi},\sigma_i}. \tag{107}$$

In addition, line 4.1 of Algorithm 1 gives that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}\widehat{Q}_{i,h}(s,\boldsymbol{a}) &= r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P\widehat{V}_{i,h+1} \\ &= r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P\widehat{V}_{i,h+1}^{\widehat{\pi},\sigma_i} = \widehat{Q}_{i,h}^{\widehat{\pi},\sigma_i}(s,\boldsymbol{a}),\end{aligned} \tag{108}$$

where the penultimate equality holds by the induction assumption and the final equality follows from (107). It indicates

$$\forall s \in \mathcal{S}: \quad \widehat{V}_{i,h}(s) = \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_h(s)}\left[\widehat{Q}_{i,h}(s,\boldsymbol{a})\right] = \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_h(s)}\left[\widehat{Q}_{i,h}^{\widehat{\pi},\sigma_i}(s,\boldsymbol{a})\right] = \widehat{V}_{i,h}^{\widehat{\pi},\sigma_i}(s) \tag{109}$$

and that the output policy obeys

$$\forall s \in \mathcal{S}: \quad \widehat{\pi}_h(\cdot \,|\, s) \leftarrow \mathsf{Compute-Nash}\left(\widehat{Q}_{1,h}^{\widehat{\pi},\sigma_i}(s,\cdot), \widehat{Q}_{2,h}^{\widehat{\pi},\sigma_i}(s,\cdot), \widehat{Q}_{n,h}^{\widehat{\pi},\sigma_i}(s,\cdot)\right). \tag{110}$$

Then the term of interest satisfies that for any $s \in \mathcal{S}$,

$$\widehat{V}_{i,h}^{\star,\widehat{\pi}_{-i},\sigma_i}(s) = \max_{\widetilde{\pi}_i:\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,h}(s)\times\widehat{\pi}_{-i,h}(s)}\left[\widehat{Q}_{i,h}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}(s,\boldsymbol{a})\right]$$

$$
\begin{aligned}
&= \max_{\widetilde{\pi}_i:\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,h}(s)\times\widehat{\pi}_{-i,h}(s)}\left[r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P\widehat{V}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,h+1}\right] \\
&\overset{\text{(i)}}{=} \max_{\widetilde{\pi}_{i,h}\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,h}(s)\times\widehat{\pi}_{-i,h}(s)}\left[r_{i,h}(s,\boldsymbol{a}) + \max_{\widetilde{\pi}_i:\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P\widehat{V}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,h+1}\right] \\
&\overset{\text{(ii)}}{=} \max_{\widetilde{\pi}_{i,h}(s)\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,h}(s)\times\widehat{\pi}_{-i,h}(s)}\left[r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})} P\widehat{V}^{\widehat{\pi},\sigma_i}_{i,h+1}\right] \\
&= \max_{\widetilde{\pi}_{i,h}(s)\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,h}(s)\times\widehat{\pi}_{-i,h}(s)}\left[\widehat{Q}^{\widehat{\pi},\sigma_i}_{i,h}(s,\boldsymbol{a})\right], \qquad (111)
\end{aligned}
$$

where (i) holds by $r_{i,h}(s,a)$ is independent from all other time steps $h' \neq h$, (ii) is due to the exchangability of $\max_{\widetilde{\pi}_i:\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)}$ and $\inf_{P\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})}$, along with the induction assumption $\widehat{V}^{\widehat{\pi},\sigma_i}_{i,h+1} = \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,h+1} = \max_{\widetilde{\pi}_i:\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} \widehat{V}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,h+1}$, and the last equality can be verified by (107). To continue, applying (110) with the definition of robust NE, one has

$$
\begin{aligned}
\widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,h}(s) &= \max_{\widetilde{\pi}_{i,h}(s)\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,h}(s)\times\widehat{\pi}_{-i,h}(s)}\left[\widehat{Q}^{\widehat{\pi},\sigma_i}_{i,h}(s,\boldsymbol{a})\right] \\
&= \mathbb{E}_{\boldsymbol{a}\in\widehat{\pi}_h(s)}\left[\widehat{Q}^{\widehat{\pi},\sigma_i}_{i,h}(s,\boldsymbol{a})\right] = \mathbb{E}_{\boldsymbol{a}\in\widehat{\pi}_h(s)}\left[\widehat{Q}_{i,h}(s,\boldsymbol{a})\right] = \widehat{V}_{i,h}(s), \qquad (112)
\end{aligned}
$$

where the penultimate equality follows from (108). Finally, it is easily observed that

$$
\forall s\in\mathcal{S}: \widehat{V}_{i,h}(s) = \mathbb{E}_{\boldsymbol{a}\in\widehat{\pi}_h(s)}\left[\widehat{Q}_{i,h}(s,\boldsymbol{a})\right] = \mathbb{E}_{\boldsymbol{a}\in\widehat{\pi}_h(s)}\left[\widehat{Q}^{\widehat{\pi},\sigma_i}_{i,h}(s,\boldsymbol{a})\right] = \widehat{V}^{\widehat{\pi},\sigma_i}_{i,h}(s). \qquad (113)
$$

Combined this fact with (115) shows that $\widehat{V}_{i,h} = \widehat{V}^{\widehat{\pi},\sigma_i}_{i,h} = \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,h}$, which complete the induction argument.

Proof for robust CCE. The proof is analogous to the above argument for robust NE. According to the different subroutine $\mathsf{Compute-CCE}$ and the corresponding output policy $\widehat{\pi}$, the proof only differs in two steps. First, for the base case, following the same routine in (106) but replacing the robust NE property by the one of robust CCE, one has

$$
\begin{aligned}
\widehat{V}_{i,H}(s) &= \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_H(s)}[\widehat{Q}_{i,H}(s,\boldsymbol{a})] = \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_H(s)}[r_{i,H}(s,\boldsymbol{a})] \\
&\geq \max_{\widetilde{\pi}_{i,H}(s)\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,H}(s)\times\widehat{\pi}_{-i,H}(s)}[r_{i,H}(s,\boldsymbol{a})] \\
&= \max_{\widetilde{\pi}_{i,H}(s)\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim\widetilde{\pi}_{i,H}(s)\times\widehat{\pi}_{-i,H}(s)}\left[\widehat{Q}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,H}(s,\boldsymbol{a})\right] \\
&= \max_{\widetilde{\pi}_i:\mathcal{S}\times[H]\to\in\Delta(\mathcal{A}_i)} \widehat{V}^{\widetilde{\pi}_i\times\widehat{\pi}_{-i},\sigma_i}_{i,H}(s) = \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,H}. \qquad (114)
\end{aligned}
$$

Secondly, following (115) in induction step, we can achieve

$$
\widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,h} \leq \widehat{V}_{i,h} \qquad (115)
$$

and $\widehat{V}_{i,h} = \widehat{V}^{\widehat{\pi},\sigma_i}_{i,h} \geq \widehat{V}^{\star,\widehat{\pi}_{-i},\sigma_i}_{i,h}$, which complete the proof.

Proof for robust CE. The proof is similar to the one of robust NE as well. According to the different subroutine $\mathsf{Compute-CE}$ and the corresponding output policy $\widehat{\pi}$, the parallel claims to (106) and (115) are shown below, which we omit the process for brevity:

$$
\begin{aligned}
\widehat{V}_{i,H}(s) &= \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_H(s)}[\widehat{Q}_{i,H}(s,\boldsymbol{a})] = \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_H(s)}[r_{i,H}(s,\boldsymbol{a})] \\
&\geq \max_{f_{i,H,s}:\mathcal{A}_i\to\mathcal{A}_i} \mathbb{E}_{\boldsymbol{a}\sim f_{i,H,s}\diamond\widehat{\pi}_H(s)}[r_{i,H}(s,\boldsymbol{a})] = \max_{f_i\in\mathcal{F}_i} \widehat{V}^{f_i\diamond\widehat{\pi},\sigma_i}_{i,H}, \qquad (116)
\end{aligned}
$$

and

$$
\max_{f_i\in\mathcal{F}_i} \widehat{V}^{f_i\diamond\widehat{\pi},\sigma_i}_{i,h} \leq \widehat{V}^{\widehat{\pi},\sigma_i}_{i,h} = \widehat{V}_{i,h}. \qquad (117)
$$

Thus we complete the proof.

C.3.2. PROOF OF LEMMA C.1

To begin with, we observe that

$$\begin{aligned}\min_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) &= \min_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi_h(s)}[Q_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a})] = \min_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi_h(s)}[r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}(P_{h,s,\boldsymbol{a}})} PV_{i,h+1}^{\pi,\sigma_i}] \\ &\geq 0 + \min_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s), \end{aligned} \tag{118}$$

where the second equality holds by the robust Bellman equation (cf. (13)). Similarly, one has

$$\begin{aligned}\max_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) &= \max_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi_h(s)}[Q_{i,h}^{\pi,\sigma_i}(s,\boldsymbol{a})] = \max_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi_h(s)}[r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}(P_{h,s,\boldsymbol{a}})} PV_{i,h+1}^{\pi,\sigma_i}] \\ &\leq 1 + \max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \inf_{P\in\mathcal{U}^{\sigma_i}(P_{h,s,\boldsymbol{a}})} PV_{i,h+1}^{\pi,\sigma_i}. \end{aligned} \tag{119}$$

Armed with above results, we are ready to prove Lemma C.1. Towards this, we introduce some additional notations for convenience. Fixing any joint policy π, note that for any $(i,h)\in[n]\times[H]$, there exist at least one state $s_{i,h}^\star$ that satisfies $V_{i,h}^{\pi,\sigma_i}(s_{i,h}^\star) = \min_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s)$.

Then, it is observed that for any $(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}$ and accessible uncertainty set $\sigma_i > 0$, we can construct an auxiliary vector $P'_{h,s,\boldsymbol{a}}\in\mathbb{R}^S$ by strictly reducing the values of some elements of $P_{h,s,\boldsymbol{a}}$ so that

$$0 \leq P'_{h,s,\boldsymbol{a}} \leq P_{h,s,\boldsymbol{a}} \quad \text{and} \quad \sum_{s'\in\mathcal{S}} P_{h,s,\boldsymbol{a}}(s') - P'_{h,s,\boldsymbol{a}}(s') = \left\|P'_{h,s,\boldsymbol{a}} - P_{h,s,\boldsymbol{a}}\right\|_1 = \sigma_i. \tag{120}$$

Recalling $e_{s_{i,h}^\star}$ denote a S-dimensional standard basis supported on $s_{i,h}^\star$, the above fact directly indicates that

$$\frac{1}{2}\left\|P'_{h,s,\boldsymbol{a}} + \sigma_i\left[e_{s_{i,h}^\star}\right]^\top - P_{h,s,\boldsymbol{a}}\right\|_1 \leq \frac{1}{2}\left\|P'_{h,s,\boldsymbol{a}} - P_{h,s,\boldsymbol{a}}\right\|_1 + \frac{1}{2}\left\|\sigma_i\left[e_{s_{i,h}^\star}\right]^\top\right\|_1 \leq \sigma_i, \tag{121}$$

where the first inequality holds by that TV distance enjoys the triangle inequality.

The above results in (121) imply that $P'_{h,s,\boldsymbol{a}} + \sigma_i\left[e_{s_{i,h}^\star}\right]^\top$ is a distribution vector and $P'_{h,s,\boldsymbol{a}} + \sigma_i\left[e_{s_{i,h}^\star}\right]^\top \in \mathcal{U}^{\sigma_i}(P_{h,s,\boldsymbol{a}})$, which leads to

$$\begin{aligned}\inf_{P\in\mathcal{U}^{\sigma_i}(P_{h,s,\boldsymbol{a}})} PV_{i,h+1}^{\pi,\sigma_i} &\leq \left(P'_{h,s,\boldsymbol{a}} + \sigma_i\left[e_{s_{i,h}^\star}\right]^\top\right) V_{i,h+1}^{\pi,\sigma_i} \leq \left\|P'_{h,s,\boldsymbol{a}}\right\|_1 \left\|V_{i,h+1}^{\pi,\sigma_i}\right\|_\infty + \sigma_i V_{i,h+1}^{\pi,\sigma_i}(s_{i,h+1}^\star) \\ &\leq (1-\sigma_i)\max_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s) + \sigma_i \min_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s), \end{aligned} \tag{122}$$

where the last inequality can be verified by (see (120))

$$\left\|P'_{h,s,\boldsymbol{a}}\right\|_1 = \sum_{s'} P'_{h,s,\boldsymbol{a}}(s') = -\sum_{s'}\left(P_{h,s,\boldsymbol{a}}(s') - P'_{h,s,\boldsymbol{a}}(s')\right) + \sum_{s'} P_{h,s,\boldsymbol{a}}(s') = 1-\sigma_i. \tag{123}$$

Inserting (122) back to (119) yields

$$\begin{aligned}\max_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) &\leq 1 + \max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \inf_{P\in\mathcal{U}^{\sigma_i}(P_{h,s,\boldsymbol{a}})} PV_{i,h+1}^{\pi,\sigma_i} \\ &\leq 1 + (1-\sigma_i)\max_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s) + \sigma_i \min_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s). \end{aligned} \tag{124}$$

Combined above fact with (118) shows that

$$\begin{aligned}\max_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) - \min_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) &\leq 1 + (1-\sigma_i)\max_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s) + \sigma_i \min_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s) - \min_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s) \\ &\leq 1 + (1-\sigma_i)\left(\max_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s) - \min_{s\in\mathcal{S}} V_{i,h+1}^{\pi,\sigma_i}(s)\right) \\ &\leq 1 + (1-\sigma_i)\left[1 + (1-\sigma_i)\left(\max_{s\in\mathcal{S}} V_{i,h+2}^{\pi}(s) - \min_{s\in\mathcal{S}} V_{i,h+2}^{\pi}(s)\right)\right] \\ &\leq \cdots \leq \frac{1-(1-\sigma_i)^{H-h}}{\sigma_i} \leq \frac{1}{\sigma_i}. \end{aligned} \tag{125}$$

Combining above result with the basic fact $\max_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) - \min_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s) \leq H-h+1$, we complete the proof.

C.3.3. PROOF OF LEMMA C.2

The proof is adapted from the routine for proving Shi et al. (2023, Lemma 9).

Step 1: a point-wise bound. Consider any fixed (independent from $\widehat{P}^0$) value vector V, combined with the definitions in (53), the $(s, \boldsymbol{a})$-th row of the term of interest can be written out as

$$\begin{aligned}
\left|P^V_{i,h,s,\boldsymbol{a}}V - \widehat{P}^V_{i,h,s,\boldsymbol{a}}V\right| &= \left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}(P^0_{h,s,\boldsymbol{a}})}\mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}(\widehat{P}^0_{h,s,\boldsymbol{a}})}\mathcal{P}V\right| \\
&\overset{\text{(i)}}{=} \Big|\max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{P^0_{h,s,\boldsymbol{a}}[V]_\alpha - \sigma_i\left(\alpha - \min_{s'}[V]_\alpha(s')\right)\right\} \\
&\quad - \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{\widehat{P}^0_{h,s,\boldsymbol{a}}[V]_\alpha - \sigma_i\left(\alpha - \min_{s'}[V]_\alpha(s')\right)\right\}\Big| \\
&\le \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left|P^0_{h,s,\boldsymbol{a}}[V]_\alpha - \widehat{P}^0_{h,s,\boldsymbol{a}}[V]_\alpha\right| \\
&\le \max_{\alpha\in[0,H]}\left|P^0_{h,s,\boldsymbol{a}}[V]_\alpha - \widehat{P}^0_{h,s,\boldsymbol{a}}[V]_\alpha\right|, \qquad (126)
\end{aligned}$$

where (i) holds by applying Lemma B.1, and the last inequality can be verified by the fact that the maximum operator is 1-Lipschitz.

To continue, recalling the definition of variance in (55) and using the Bernstein's inequality, one has for a fixed $\alpha \in [0, H]$ and $(s, \boldsymbol{a}) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$,

$$\begin{aligned}
\left|\left(P^0_{h,s,\boldsymbol{a}} - \widehat{P}^0_{h,s,\boldsymbol{a}}\right)[V]_\alpha\right| &\le \sqrt{\frac{2\log(\frac{2}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}([V]_\alpha)} + \frac{2H\log(\frac{2}{\delta})}{3N} \\
&\le \sqrt{\frac{2\log(\frac{2}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}(V)} + \frac{2H\log(\frac{2}{\delta})}{3N}, \qquad (127)
\end{aligned}$$

where the first inequality holds by the fact that $\|V\|_\infty \le H$, and the last inequality can be easily verified by noticing that $\mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}([V]_\alpha) \le \mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}(V)$ for all $\alpha \in [0, \max_s V(s)]$.

Step 2: the union bound. Then to obtain the union bound, we first notice that the function $\left|\left(P^0_{h,s,\boldsymbol{a}} - \widehat{P}^0_{h,s,\boldsymbol{a}}\right)[V]_\alpha\right|$ is 1-Lipschitz w.r.t. α for any V obeying $0 \le V(s) \le H$. Therefore, we can construct an ε_1-net N_{ε_1} for α over $[0, H]$ with the size up to $|N_{\varepsilon_1}| \le \frac{3H}{\varepsilon_1}$ (Vershynin, 2018). So applying the uniform concentration argument combined with (127) yields that for all $(\alpha, s, \boldsymbol{a}) \in N_{\varepsilon_1}\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$,

$$\left|\left(P^0_{h,s,\boldsymbol{a}} - \widehat{P}^0_{h,s,\boldsymbol{a}}\right)[V]_\alpha\right| \le \sqrt{\frac{2\log\left(\frac{2S\prod_{i=1}^n A_i|N_{\varepsilon_1}|}{\delta}\right)}{N}}\sqrt{\mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}(V)} + \frac{2H\log\left(\frac{2S\prod_{i=1}^n A_i|N_{\varepsilon_1}|}{\delta}\right)}{3N}. \qquad (128)$$

Inserting the above fact back to (126), we arrive at: for all $(s, \boldsymbol{a}) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\left|P^V_{i,h,s,\boldsymbol{a}}V - \widehat{P}^V_{i,h,s,\boldsymbol{a}}V\right| &\le \max_{\alpha\in[0,H]}\left|P^0_{h,s,\boldsymbol{a}}[V]_\alpha - \widehat{P}^0_{h,s,\boldsymbol{a}}[V]_\alpha\right| \\
&\overset{\text{(i)}}{\le} \sup_{\alpha\in N_{\varepsilon_1}}\left|P^0_{h,s,\boldsymbol{a}}[V]_\alpha - \widehat{P}^0_{h,s,\boldsymbol{a}}[V]_\alpha\right| + \varepsilon_1 \\
&\overset{\text{(ii)}}{\le} \sqrt{\frac{2\log\left(\frac{2S\prod_{i=1}^n A_i|N_{\varepsilon_1}|}{\delta}\right)}{N}}\sqrt{\mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}(V)} + \frac{2\log\left(\frac{2S\prod_{i=1}^n A_i|N_{\varepsilon_1}|}{\delta}\right)H}{3N} + \varepsilon_1 \qquad (129) \\
&\overset{\text{(iii)}}{\le} \sqrt{\frac{2\log\left(\frac{2S\prod_{i=1}^n A_i|N_{\varepsilon_1}|}{\delta}\right)}{N}}\sqrt{\mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}(V)} + \frac{\log\left(\frac{2S\prod_{i=1}^n A_i|N_{\varepsilon_1}|}{\delta}\right)H}{N}
\end{aligned}$$

$$\overset{\text{(iv)}}{\leq} 2\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_iN}{\delta}\right)}{N}}\sqrt{\mathrm{Var}_{P^0_{h,s,\boldsymbol{a}}}(V)}+\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_iN}{\delta}\right)H}{N} \tag{130}$$

$$\leq 2\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_iN}{\delta}\right)}{N}}\|V\|_\infty+\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_iN}{\delta}\right)H}{N}$$

$$\leq 3\sqrt{\frac{H^2\log\left(\frac{18S\prod_{i=1}^{n}A_iN}{\delta}\right)}{N}} \tag{131}$$

where (i) arises from the fact that the solution $\alpha^\star=\arg\max_{\alpha\in[0,H]}\left|P^0_{h,s,\boldsymbol{a}}[V]_\alpha-\widehat{P}^0_{h,s,\boldsymbol{a}}[V]_\alpha\right|$ falls into the ε_1-ball centered around some point inside N_{ε_1} and $\left|P^0_{h,s,\boldsymbol{a}}[V]_\alpha-\widehat{P}^0_{h,s,\boldsymbol{a}}[V]_\alpha\right|$ is 1-Lipschitz w.r.t. α, (ii) holds by (128), (iii) follows from taking $\varepsilon_1=\frac{\log(\frac{2S\prod_{i=1}^{n}A_i|N_{\varepsilon_1}|}{\delta})H}{3N}$, (iv) is verified by $|N_{\varepsilon_1}|\leq\frac{3H}{\varepsilon_1}\leq 9N$, and the last inequality is due to the fact $\|V\|_\infty\leq H$ and letting $N\geq\log(\frac{18S\prod_{i=1}^{n}A_iN}{\delta})$.

Invoking the matrix form (see (53) and (54)) and applying the above result with $V=\widehat{V}^{\pi,\sigma_i}_{i,h+1}$ for a union bound over all $(h,i,s,\boldsymbol{a})\in[H]\times[n]\times\mathcal{S}\times\mathcal{A}$, we complete the proof: with probability at least $1-\delta$,

$$\begin{aligned}
\forall(h,i)\in[H]\times[n]:\quad a^\pi_{i,h}&=\left|\underline{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1}-\underline{\widehat{P}}^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1}\right|\\
&=\left|\Pi^\pi_hP^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1}-\Pi^\pi_h\widehat{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1}\right|\\
&\overset{\text{(i)}}{\leq}\Pi^\pi_h\left|P^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1}-\widehat{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1}\right| \qquad (132)\\
&\leq 2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\Pi^\pi_h\sqrt{\mathsf{Var}_{P^0_h}(\widehat{V}^\pi_{i,h+1})}+\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})H}{N}1\\
&\overset{\text{(ii)}}{\leq} 2\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}\sqrt{\mathsf{Var}_{\underline{P}^\pi_h}(\widehat{V}^\pi_{i,h+1})}+\frac{\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})H}{N}1\\
&\leq 3\sqrt{\frac{H^2\log(\frac{18S\prod_{i=1}^{n}A_inHN}{\delta})}{N}}1, \qquad (133)
\end{aligned}$$

where (i) and (ii) hold by the Jensen's inequality, $\mathsf{Var}(\cdot)$ is defined in (56), and $P^0_h,\underline{P}^\pi_h$ are defined in (54).

C.3.4. Proof of Lemma C.3

In this section, we want to take the accessible range of the robust value function $\widehat{V}^{\pi,\sigma_i}_{i,j+1}$ into consideration when controlling $\sum_{j=h}^{H}\left\langle d^j_h,\mathsf{Var}_{\underline{P}^{\pi,\widehat{V}}_{i,j}}(\widehat{V}^{\pi,\sigma_i}_{i,j+1})\right\rangle$. Towards this, we introduce some auxiliary values and reward functions as below. For any time step $h\in[H]$ and the i-th agent:

- $\widehat{V}^{\min}_h:=\min_{s\in\mathcal{S}}\widehat{V}^{\pi,\sigma_i}_{i,h}(s)$: $\widehat{V}^{\min}_h$ denote the minimum value of all the entries in vector $\widehat{V}^{\pi,\sigma_i}_{i,h}$.
- $\widehat{V}'_h:=\widehat{V}^{\pi,\sigma_i}_{i,h}-\widehat{V}^{\min}_h1$: truncated value function.
- $\widehat{r}^{\min}_{i,h}=r^\pi_{i,h}+\left(\widehat{V}^{\min}_{h+1}-\widehat{V}^{\min}_h\right)1$: truncated reward function.

With above notations, we introduce the following fact of V'_h:

$$\widehat{V}'_h=\widehat{V}^{\pi,\sigma_i}_{i,h}-\widehat{V}^{\min}_h1\overset{\text{(i)}}{=}r^\pi_{i,h}+\underline{\widehat{P}}^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1}-\widehat{V}^{\min}_h1$$

$$
\begin{aligned}
&= r^{\pi}_{i,h} + \underline{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}^{\pi,\sigma_i}_{i,h+1} + \left(\widehat{\underline{P}}^{\pi,\widehat{V}}_{i,h} - \underline{P}^{\pi,\widehat{V}}_{i,h}\right)\widehat{V}^{\pi,\sigma_i}_{i,h+1} - \widehat{V}^{\min}_{h}1 \\
&= r^{\pi}_{i,h} + \left(\widehat{V}^{\min}_{h+1} - \widehat{V}^{\min}_{h}\right)1 + \underline{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}'_{h+1} + \left(\widehat{\underline{P}}^{\pi,\widehat{V}}_{i,h} - \underline{P}^{\pi,\widehat{V}}_{i,h}\right)\widehat{V}^{\pi,\sigma_i}_{i,h+1} \\
&= \widehat{r}^{\min}_{i,h} + \underline{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}'_{h+1} + \left(\widehat{\underline{P}}^{\pi,\widehat{V}}_{i,h} - \underline{P}^{\pi,\widehat{V}}_{i,h}\right)\widehat{V}^{\pi,\sigma_i}_{i,h+1}, \qquad (134)
\end{aligned}
$$

where (i) holds by the robust Bellman's consistency equation in (62).

With the above fact in hand, we can verify that

$$
\begin{aligned}
\mathsf{Var}_{\underline{P}^{\pi,\widehat{V}}_{i,h}}(\widehat{V}^{\pi,\sigma_i}_{i,h+1}) &\overset{\text{(i)}}{=} \mathrm{Var}_{\underline{P}^{\pi,\widehat{V}}_{i,h}}(\widehat{V}'_{h+1}) = \underline{P}^{\pi,\widehat{V}}_{i,h}\left(\widehat{V}'_{h+1}\circ\widehat{V}'_{h+1}\right) - \left(\underline{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}'_{h+1}\right)\circ\left(\underline{P}^{\pi,\widehat{V}}_{i,h}\widehat{V}'_{h+1}\right) \\
&\overset{\text{(ii)}}{=} \underline{P}^{\pi,\widehat{V}}_{i,h}\left(\widehat{V}'_{h+1}\circ\widehat{V}'_{h+1}\right) - \left(\widehat{V}'_{h} - \widehat{r}^{\min}_{i,h} - \left(\widehat{\underline{P}}^{\pi,\widehat{V}}_{i,h} - \underline{P}^{\pi,\widehat{V}}_{i,h}\right)\widehat{V}^{\pi,\sigma_i}_{i,h+1}\right)^{\circ 2} \\
&= \underline{P}^{\pi,\widehat{V}}_{i,h}\left(\widehat{V}'_{h+1}\circ\widehat{V}'_{h+1}\right) - \widehat{V}'_{h}\circ\widehat{V}'_{h} + 2\widehat{V}'_{h}\circ\left(\widehat{r}^{\min}_{i,h} + \left(\widehat{\underline{P}}^{\pi,\widehat{V}}_{i,h} - \underline{P}^{\pi,\widehat{V}}_{i,h}\right)\widehat{V}^{\pi,\sigma_i}_{i,h+1}\right) \\
&\quad - \left(\widehat{r}^{\min}_{i,h} + \left(\widehat{\underline{P}}^{\pi,\widehat{V}}_{i,h} - \underline{P}^{\pi,\widehat{V}}_{i,h}\right)\widehat{V}^{\pi,\sigma_i}_{i,h+1}\right)^{\circ 2} \\
&\overset{\text{(iii)}}{\leq} \underline{P}^{\pi,\widehat{V}}_{i,h}\left(\widehat{V}'_{h+1}\circ\widehat{V}'_{h+1}\right) - \widehat{V}'_{h}\circ\widehat{V}'_{h} + 2\left\|\widehat{V}'_{h}\right\|_{\infty}\left(1 + \left|\left(\widehat{\underline{P}}^{\pi,\widehat{V}}_{i,h} - \underline{P}^{\pi,\widehat{V}}_{i,h}\right)\widehat{V}^{\pi,\sigma_i}_{i,h+1}\right|\right) \qquad (135) \\
&\leq \underline{P}^{\pi,\widehat{V}}_{i,h}\left(\widehat{V}'_{h+1}\circ\widehat{V}'_{h+1}\right) - \widehat{V}'_{h}\circ\widehat{V}'_{h} + 2\left\|\widehat{V}'_{h}\right\|_{\infty}1 + 6\|V'_{h}\|_{\infty}\sqrt{\frac{H^2\log\left(\frac{18S\prod_{i=1}^{n}A_i nHN}{\delta}\right)}{N}}1 \qquad (136)
\end{aligned}
$$

holds with probability at least $1-\delta$, where (i) follows from the fact that $\mathrm{Var}_{\underline{P}^{\pi,\widehat{V}}_{i,h}}(V - b1) = \mathrm{Var}_{\underline{P}^{\pi,\widehat{V}}_{i,h}}(V)$ for any value vector $V \in \mathbb{R}^S$ and scalar b, (ii) holds by (134), (iii) arises from $\widehat{r}^{\min}_{i,h} \leq r^{\pi}_{i,h} \leq 1$ since $V^{\min}_{h+1} - V^{\min}_{h} \leq 0$ by definition, and the last inequality holds by (133).

Finally, combining (136) and the definition of d^j_h in (77), the term of interest can be controlled as

$$
\begin{aligned}
&\sum_{j=h}^{H}\left\langle d^j_h, \mathsf{Var}_{\underline{P}^{\pi,\widehat{V}}_{i,j}}(\widehat{V}^{\pi,\sigma_i}_{i,j+1})\right\rangle \\
&= \sum_{j=h}^{H}(d^j_h)^{\top}\left(\underline{P}^{\pi,\widehat{V}}_{i,j}\left(\widehat{V}'_{j+1}\circ\widehat{V}'_{j+1}\right) - \widehat{V}'_{j}\circ\widehat{V}'_{j} + 2\|\widehat{V}'_{j}\|_{\infty}1 + 6\|\widehat{V}'_{j}\|_{\infty}\sqrt{\frac{H^2\log\left(\frac{18S\prod_{i=1}^{n}A_i nHN}{\delta}\right)}{N}}1\right) \\
&\overset{\text{(i)}}{\leq} \sum_{j=h}^{H}\left[(d^j_h)^{\top}\left(\underline{P}^{\pi,\widehat{V}}_{i,j}\left(\widehat{V}'_{j+1}\circ\widehat{V}'_{j+1}\right) - \widehat{V}'_{j}\circ\widehat{V}'_{j}\right)\right] + 2H\|\widehat{V}'_{h}\|_{\infty} + 6H^2\|\widehat{V}'_{h}\|_{\infty}\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_i nHN}{\delta}\right)}{N}} \\
&= \sum_{j=h}^{H}\left[(d^{j+1}_h)^{\top}\left(\widehat{V}'_{j+1}\circ\widehat{V}'_{j+1}\right) - (d^j_h)^{\top}\left(\widehat{V}'_{j}\circ\widehat{V}'_{j}\right)\right] + 2H\|\widehat{V}'_{h}\|_{\infty} + 6H^2\|\widehat{V}'_{h}\|_{\infty}\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_i nHN}{\delta}\right)}{N}} \\
&\leq \left\|d^{H+1}_h\right\|_1\left\|\widehat{V}'_{H+1}\circ\widehat{V}'_{H+1}\right\|_{\infty} + 2H\|\widehat{V}'_{h}\|_{\infty} + 6H^2\|\widehat{V}'_{h}\|_{\infty}\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_i nHN}{\delta}\right)}{N}} \\
&\leq 3H\|\widehat{V}'_{h}\|_{\infty} + 6H^2\|\widehat{V}'_{h}\|_{\infty}\sqrt{\frac{\log\left(\frac{18S\prod_{i=1}^{n}A_i nHN}{\delta}\right)}{N}} \\
&= 3H\|\widehat{V}'_{h}\|_{\infty}\left(1 + 2H\sqrt{\frac{\log(\frac{18S\prod_{i=1}^{n}A_i nHN}{\delta})}{N}}\right), \qquad (137)
\end{aligned}
$$

where (i) holds by the fact $\|\widehat{V}_h'\|_\infty \geq \|\widehat{V}_{h+1}'\|_\infty \geq \cdots \geq \|\widehat{V}_H'\|_\infty$ and basic calculus.

C.3.5. Proof of Lemma C.4

We start with the proof about the empirical MG $\mathcal{MG}_{\mathsf{rob}}$. To begin with, for any policy π and the i-th agent, we define

$$\forall h \in [H]: \quad V_{i,h}^{\mathsf{span}} \coloneqq \widehat{V}_{i,h}^{\pi,\sigma_i} - \min_{s' \in \mathcal{S}} \widehat{V}_{i,h}^{\pi,\sigma_i}(s')1, \tag{138}$$

which leads to

$$\left\| V_{i,h}^{\mathsf{span}} \right\|_\infty \leq \min\left\{ \frac{1}{\sigma_i}, H-h+1 \right\}. \tag{139}$$

which holds by applying Lemma C.1.

Armed with above notation and facts, considering any transition kernel $P' \in \mathbb{R}^S$ and any $\widetilde{P} \in \mathbb{R}^S$ obeying $\widetilde{P} \in \mathcal{U}^{\sigma_i}(P')$, we have for all $(i,h) \in [n] \times [H]$

$$\begin{aligned}
\left| \mathsf{Var}_{P'}(\widehat{V}_{i,h}^{\pi,\sigma_i}) - \mathsf{Var}_{\widetilde{P}}(\widehat{V}_{i,h}^{\pi,\sigma_i}) \right| &= \left| \mathsf{Var}_{P'}(V_{i,h}^{\mathsf{span}}) - \mathsf{Var}_{\widetilde{P}}(V_{i,h}^{\mathsf{span}}) \right| \\
&\leq \left\| \widetilde{P} - P' \right\|_1 \left\| V_{i,h}^{\mathsf{span}} \right\|_\infty \\
&\leq \sigma_i \left(\min\left\{ \frac{1}{\sigma_i}, H-h+1 \right\} \right)^2 \leq \min\left\{ \frac{1}{\sigma_i}, H-h+1 \right\}.
\end{aligned} \tag{140}$$

Similar facts can be verified for standard MG $\mathcal{MG}$ analogously.

C.3.6. Proof of Lemma C.5

Analogous to Appendix C.3.4, we introduce some auxiliary values and reward functions to control

$$\sum_{j=h}^{H} \left\langle w_h^j, \mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}(V_{i,j+1}^{\pi,\sigma_i}) \right\rangle$$

as below: for any time step h and the i-th agent

- $V_h^{\min} \coloneqq \min_{s \in \mathcal{S}} V_{i,h}^{\pi,\sigma_i}(s)$: $V_h^{\min}$ denote the minimum value of all the entries in vector $V_{i,h}^{\pi,\sigma_i}$.
- $V_h' \coloneqq V_{i,h}^{\pi,\sigma_i} - V_h^{\min}1$: truncated value function.
- $r_{i,h}^{\min} = r_{i,h}^{\pi} + \left(V_{h+1}^{\min} - V_h^{\min}\right)1$: truncated reward function.

Then applying the robust Bellman's consistency equation in (59) gives

$$\begin{aligned}
V_h' = V_{i,h}^{\pi,\sigma_i} - V_h^{\min}1 &= r_{i,h}^{\pi} + \underline{P}_{i,h}^{\pi,V} V_{i,h+1}^{\pi,\sigma_i} - V_h^{\min}1 \\
&= r_{i,h}^{\pi} + \left(V_{h+1}^{\min} - V_h^{\min}\right)1 + \underline{P}_{i,h}^{\pi,V} V_{h+1}' = r_{i,h}^{\min} + \underline{P}_{i,h}^{\pi,V} V_{h+1}'.
\end{aligned} \tag{141}$$

The above fact leads to

$$\begin{aligned}
\mathsf{Var}_{\underline{P}_{i,h}^{\pi,V}}(V_{i,h+1}^{\pi,\sigma_i}) &\overset{\text{(i)}}{=} \mathrm{Var}_{\underline{P}_{i,h}^{\pi,V}}(V_{h+1}') = \underline{P}_{i,h}^{\pi,V}\left(V_{h+1}' \circ V_{h+1}'\right) - \left(\underline{P}_{i,h}^{\pi,V} V_{h+1}'\right) \circ \left(\underline{P}_{i,h}^{\pi,V} V_{h+1}'\right) \\
&\overset{\text{(ii)}}{=} \underline{P}_{i,h}^{\pi,V}\left(V_{h+1}' \circ V_{h+1}'\right) - \left(V_h' - r_{i,h}^{\min}\right)^{\circ 2} \\
&= \underline{P}_{i,h}^{\pi,V}\left(V_{h+1}' \circ V_{h+1}'\right) - V_h' \circ V_h' + 2V_h' \circ r_{i,h}^{\min} - r_{i,h}^{\min} \circ r_{i,h}^{\min} \\
&\leq \underline{P}_{i,h}^{\pi,V}\left(V_{h+1}' \circ V_{h+1}'\right) - V_h' \circ V_h' + 2\|V_h'\|_\infty 1
\end{aligned} \tag{142}$$

where (i) follows from the fact that $\mathrm{Var}_{\underline{P}_{i,h}^{\pi,V}}(V-b1)=\mathrm{Var}_{\underline{P}_{i,h}^{\pi,\hat{V}}}(V)$ for any value vector $V\in\mathbb{R}^S$ and scalar b, (ii) holds by (141), and the last inequality arises from $r_{i,h}^{\min}\le r_{i,h}^{\pi}\le 1$ since $V_{h+1}^{\min}-V_h^{\min}\le 0$ by definition.

Consequently, combining (142) and the definition of w_h^j in (85), we arrive at

$$\begin{aligned}
&\sum_{j=h}^{H}\left\langle w_h^j,\mathsf{Var}_{\underline{P}_{i,j}^{\pi,V}}\left(V_{i,j+1}^{\pi,\sigma_i}\right)\right\rangle\\
&=\sum_{j=h}^{H}(w_h^j)^\top\left(\underline{P}_{i,j}^{\pi,V}\left(V'_{j+1}\circ V'_{j+1}\right)-V'_j\circ V'_j+2\|V'_h\|_\infty 1\right)\\
&\overset{\text{(i)}}{\le}\sum_{j=h}^{H}\left[(w_h^j)^\top\left(\underline{P}_{i,j}^{\pi,V}\left(V'_{j+1}\circ V'_{j+1}\right)-V'_j\circ V'_j\right)\right]+2H\|V'_h\|_\infty\\
&=\sum_{j=h}^{H}\left[(w_h^{j+1})^\top\left(V'_{j+1}\circ V'_{j+1}\right)-(w_h^j)^\top\left(V'_j\circ V'_j\right)\right]+2H\|V'_h\|_\infty\\
&\le\|w_h^{H+1}\|_1\left\|V'_{H+1}\circ V'_{H+1}\right\|_\infty+2H\|V'_h\|_\infty\\
&\le 3H\|V'_h\|_\infty,
\end{aligned}\tag{143}$$

where (i) and the last inequality hold by the fact $\|V'_h\|_\infty\ge\|V'_{h+1}\|_\infty\ge\cdots\ge\|V'_H\|_\infty$ and basic calculus.

D. Proof of Theorem 4.2

In this section, the proof will focus on a special and simpler class of RMGs: distributionally robust Markov decision processes (RMDPs) — single-agent RMGs.

Before proceeding, to keep self-contained, we first briefly introduce the definition of a RMDP in finite-horizon episodic setting. Recall that a multi-agent general-sum robust Markov games (RMG) with TV uncertainty set can be represented as $\mathcal{MG}=\left\{\mathcal{S},\{\mathcal{A}_i\}_{1\le i\le n},\{\mathcal{U}^{\sigma_i}(P^0)\}_{1\le i\le n},r,H\right\}$. Resorting to the same notations for RMGs, a finite-horizon episodic distributionally robust MDP (RMDP) can be represented as $\mathcal{M}_{\mathsf{rob}}=\left(\mathcal{S},\mathcal{A}_1,\mathcal{U}^{\sigma_1}(P^0),\{r_{1,h}\}_{1\le h\le H},H\right)$, i.e., let $n=1$. Then we can show an essential fact between RMGs and RMDPs that allow us to turn to RMDPs for proving Theorem 4.2. Without loss of generality, we consider the class of RMGs with n players that obey $|\mathcal{A}_1|\ge\max\{|\mathcal{A}_2|,\cdots,|\mathcal{A}_m|\}$. Moreover, let $|\mathcal{A}_2|=|\mathcal{A}_3|=\cdots=|\mathcal{A}_m|=1$ for simplicity, which leaves those agents' ($i=2,3,\cdots,n$) choices of actions having no randomness or effects on the transitions or rewards for any agents. Consequently, it is clear that finding a robust NE/CE/CCE of such RMGs degrades to finding the optimal policy of the first agent over a corresponding RMDP $\mathcal{M}_{\mathsf{rob}}=\left\{\mathcal{S},\mathcal{A}_1,\mathcal{U}^{\sigma_1}(P^0),\{r_{1,h}\}_{1\le h\le H},H\right\}$.

Therefore, in this section, we turn to construct the lower bound for finding the optimal policy over RMDPs instead, which directly imply a lower bound for finding equilibriums (robust NE/CE/CCE) of RMGs.

Before continuing, we make note of the following useful property about the KL divergence in Tsybakov (2009, Lemma 2.7) which is useful in this section.

Lemma D.1. *For any $p,q\in(0,1)$, it holds that*

$$\mathsf{KL}(p\parallel q)\le\frac{(p-q)^2}{q(1-q)}.\tag{144}$$

D.1. Constructing hard robust MDP instances

The hard instances developed here are different from standard MDP since we need to consider that the transition kernel can be perturbed in robust MDPs. This is the first lower bound for robust MDPs in episodic setting.

Step 1: constructing hard robust MDP instances. To begin with, we first introduce an auxiliary collection $\Theta\subseteq\{0,1\}^H$, consisting of H-dimensional vectors. In addition, resorting to the Gilbert-Varshamov lemma (Gilbert, 1952), we notice that

there exists a set $\Theta \subseteq \{0,1\}^H$ such that:

$$\text{for any } \theta, \widetilde{\theta} \in \Theta \text{ obeying } \theta \neq \widetilde{\theta}: \quad \|\theta - \widetilde{\theta}\|_1 \geq \frac{H}{8} \qquad \text{and} \qquad |\Theta| \geq e^{H/8}. \tag{145}$$

Without loss of generality, we denote the first component of Θ as θ^{base} and denote $\Theta^\star$ as $\Theta \setminus \{\theta^{\mathsf{base}}\}$. With this in mind, we construct a set of RMDPs as below:

$$\mathcal{M}(\mathcal{W}, \Theta) \coloneqq \left\{ \mathcal{M}_w^\theta = \left(\mathcal{S}, \mathcal{A}, \mathcal{U}^\sigma(P^{w,\theta}), \{r_h\}_{h=1}^H, H\right) \mid w \in \mathcal{W} = \{0, 1, \cdots, SA-1\}, \theta = [\theta_h]_{1 \leq h \leq H} \in \Theta^\star \right\}, \tag{146}$$

where

$$\mathcal{S} = \{0, 1, \ldots, S-1\}, \qquad \text{and} \qquad \mathcal{A} = \{0, 1, \cdots, A-1\},$$

and σ will be introduced momentarily.

In words, the collection of $\mathcal{M}(\mathcal{W}, \Theta)$ consists of $|\mathcal{W}| = SA$ subsets, with each includes $|\Theta^\star|$ different RMDPs associated with some $w \in \mathcal{W}$. The state space of each RMDP $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W}, \Theta)$ is denoted as $\mathcal{S}_\mathcal{M}$, includes two classes of states $\mathcal{X} = \{x_i \mid i \in \mathcal{W}\}$ and $\mathcal{Y} = \{y_i \mid i \in \mathcal{W}\}$. Each state in $\mathcal{X}$ and $\mathcal{Y}$ only have two possible actions $\mathcal{A}_\mathcal{M} = \{0, 1\}$. So we have totally $2|\mathcal{W}| = 2SA$ states and there is in total $|\mathcal{S}_\mathcal{M}||\mathcal{A}_\mathcal{M}| = 4SA$ state-action pairs.

We shall define the nominal transition kernels for $\mathcal{M}(\mathcal{W}, \Theta)$, where any state $x_i \in \mathcal{X}$ only transits to the corresponding $y_i \in \mathcal{Y}$ or itself. For convenience, for any $s = x_i \in \mathcal{X}$, we denote the corresponding state $y_i \in \mathcal{Y}$ as $s^{x \to y}$.

Armed with above notations, we define a basic nominal transition kernel associated with θ^{base} as below: for all $(h, s, a) \in [H] \times \mathcal{S}_\mathcal{M} \times \mathcal{A}_\mathcal{M}$,

$$P_h^\star(s' \mid s, a) = \begin{cases} (p+\Delta)\mathbb{1}(s' = s^{x \to y}) + (1-p-\Delta)\mathbb{1}(s' = s) & \text{if} \quad s \in \mathcal{X}, a = \theta_h^{\mathsf{base}} \\ p\mathbb{1}(s' = s^{x \to y}) + (1-p)\mathbb{1}(s' = s) & \text{if} \quad s \in \mathcal{X}, a = 1 - \theta_h^{\mathsf{base}} \\ \mathbb{1}(s' = s) & \text{if} \quad s \in \mathcal{Y}. \end{cases} \tag{147}$$

In addition, for any RMDP $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W}, \Theta)$, the transition kernel $P^{w,\theta} = \{P_h^{w,\theta}\}_{h=1}^H$ is specified as follows: for any $(s, a, s', h) \in \mathcal{S}_\mathcal{M} \times \mathcal{A}_\mathcal{M} \times \mathcal{S}_\mathcal{M} \times [H]$,

$$P_h^{w,\theta}(s' \mid s, a) = \begin{cases} p\mathbb{1}(s' = y_w) + (1-p)\mathbb{1}(s' = s) & \text{if} \quad s = x_w, a = \theta_h \\ q\mathbb{1}(s' = y_w) + (1-q)\mathbb{1}(s' = s) & \text{if} \quad s = x_w, a = 1 - \theta_h \\ P_h^\star(s' \mid s, a) & \text{otherwise} \end{cases} \tag{148}$$

Here, p and q are set according to

$$0 \leq p \leq p + \Delta \leq 1 \quad \text{and} \quad 0 \leq q = p - \Delta \tag{149}$$

for some p and $\Delta > 0$ that will be introduced momentarily. In words, the transition kernel of each $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W}, \Theta)$ only differs slightly from the basic nominal transition kernel $P_h^\star$ when $s = x_w$, which makes all the components within $\mathcal{M}(\mathcal{W}, \Theta)$ closed to each other.

To continue, the reward function is defined as

$$\forall (h, s, a) \in [H] \times \mathcal{S}_\mathcal{M} \times \{0, 1\}: \quad r_h(s, a) = \begin{cases} 1 & \text{if } s \in \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases} \tag{150}$$

Uncertainty set of the transition kernels. Denote the transition kernel vector as

$$\forall (h, s, a) \in [H] \times \mathcal{S}_\mathcal{M} \times \{0, 1\}: \quad P_{h,s,a}^{w,\theta} \coloneqq P_h^{w,\theta}(\cdot \mid s, a) \in \Delta(\mathcal{S}). \tag{151}$$

Recalling the uncertainty set defined in (8), we know $\mathcal{U}^\sigma(P^{w,\theta})$ represents:

$$\mathcal{U}^\sigma(P^{w,\theta}) \coloneqq \otimes \mathcal{U}^\sigma(P_{h,s,a}^{w,\theta}), \qquad \mathcal{U}^\sigma(P_{h,s,a}^{w,\theta}) \coloneqq \left\{ \widetilde{P}_{h,s,a}^{w,\theta} \in \Delta(\mathcal{S}) : \frac{1}{2}\left\| \widetilde{P}_{h,s,a}^{w,\theta} - P_{h,s,a}^{w,\theta} \right\|_1 \leq \sigma \right\}, \tag{152}$$

where $\otimes$ represents the Cartesian product over $(h,s,a) \in [H] \times \mathcal{S}_{\mathcal{M}} \times \mathcal{A}_{\mathcal{M}}$.

For such TV uncertainty set, without loss of generality, let the uncertainty level to be $\sigma \in (0, 1-c_0]$ for some $0 < c_0 < 1$. Then taking $c_2 \le \frac{1}{4}$ amd $c_1 \coloneqq \frac{c_0}{2} \le \frac{1}{4}$, p and Δ are set as

$$p = \begin{cases} \frac{c_2}{H}, & \text{if } \sigma \le \frac{c_2}{2H} \\ \left(1+\frac{c_1}{H}\right)\sigma & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Delta \le \begin{cases} \frac{c_2}{2H}, & \text{if } \sigma \le \frac{c_2}{2H} \\ \frac{c_1}{H}\sigma & \text{otherwise} \end{cases} \tag{153}$$

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

$$\begin{aligned} &\text{when } \sigma > \frac{c_2}{2H}: \quad \left(1+\frac{c_1}{H}\right)\sigma + \frac{c_1}{H}\sigma \le 1 - c_0 + \frac{2c_1}{H}\sigma \le 1 - \frac{c_0(H-1)}{H} < 1, \\ &\text{when } \sigma \le \frac{c_2}{2H}: \quad \frac{3c_2}{2H} \le 1. \end{aligned} \tag{154}$$

Then we introduce some useful notations and facts throughout this section. First, for any RMDP $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W}, \Theta)$ and any $(h,s,a,s') \in [H] \times \mathcal{S}_{\mathcal{M}} \times \mathcal{A}_{\mathcal{M}} \times \mathcal{S}_{\mathcal{M}}$, we denote the minimum probability of transiting from (s,a) to s' determined by any perturbed transition kernel $P_{h,s,a} \in \mathcal{U}^\sigma(P_{h,s,a}^{w,\theta})$ as

$$\underline{P}_h^{w,\theta}(s' \,|\, s,a) \coloneqq \inf_{P_{h,s,a} \in \mathcal{U}^\sigma(P_{h,s,a}^{w,\theta})} P_h(s' \,|\, s,a) = \max\{P_h(s' \,|\, s,a) - \sigma, 0\}, \tag{155}$$

where the last equation can be easily verified by the definition of $\mathcal{U}^\sigma(\cdot)$ in (152) and distributing the probability on s' to other states.

Especially, for convenience, we denote the transition from each $s \in \mathcal{X}$ to the corresponding state $s^{x \to y} \in \mathcal{Y}$ of any $\mathcal{M}_w^\theta$ as below, which plays an important role in the analysis: for all $h \in [H]$,

$$\begin{aligned} \text{for } x_w: \quad & \underline{p}_h \coloneqq \underline{P}_h^{w,\theta}(y_w \,|\, x_w, \theta_h) = p - \sigma, \qquad \underline{q}_h \coloneqq \underline{P}_h^{w,\theta}(y_w \,|\, x_w, 1-\theta_h) = q - \sigma, \\ \text{for } s \in \mathcal{X} \setminus \{x_w\}: \quad & \underline{p}'_h \coloneqq \underline{P}_h^{w,\theta}(s^{x\to y} \,|\, s, \theta_h^{\mathsf{base}}) = p + \Delta - \sigma, \qquad \underline{q}'_h \coloneqq \underline{P}_h^{w,\theta}(s^{x\to y} \,|\, s, 1-\theta_h^{\mathsf{base}}) = p - \sigma, \end{aligned} \tag{156}$$

which follows from the following fact that is clear from (153)

$$p + \Delta \ge p \ge q = p - \Delta \ge \max\left\{\frac{c_2}{2H}, \sigma\right\}. \tag{157}$$

Then it is obvious that

$$\underline{p}_1 = \underline{p}_2 = \cdots \underline{p}_H, \quad \underline{q}_1 = \underline{q}_2 = \cdots \underline{q}_H, \quad \underline{p}'_1 = \underline{p}'_2 = \cdots \underline{p}'_H, \quad \underline{q}'_1 = \underline{q}'_2 = \cdots \underline{q}'_H, \tag{158}$$

which motivates us to abbreviate them consistently as $\underline{p} \coloneqq \underline{p}_1$, $\underline{q} \coloneqq \underline{q}_1$, $\underline{p}' \coloneqq \underline{p}'_1$, and $\underline{q}' \coloneqq \underline{q}'_1$ later.

Robust value functions and optimal policies. Now we are ready to characterize the corresponding robust value functions and identify the optimal policies for RMDP instances. With abuse of notations, for any RMDP $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W}, \Theta)$, we denote $\pi^{\star,w,\theta} = \{\pi_h^{\star,w,\theta}\}_{h=1}^H$ as the optimal policy. In addition, at each step h, we let $V_h^{\pi,\sigma,w,\theta}$ (resp. $V_h^{\star,\sigma,w,\theta}$) represent the robust value function of any policy π (resp. $\pi^{\star,w,\theta}$) with uncertainty level σ. Armed with these notations, the following lemma shows some essential properties concerning the robust value functions and optimal policies; the proof is postponed to Appendix D.3.1.

Lemma D.2. *Consider any $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W}, \Theta)$ and any policy π. Defining*

$$x_h^{\pi,w,\theta} = \underline{p}\pi_h(\theta_h \,|\, x_w) + \underline{q}\pi_h(1-\theta_h \,|\, x_w), \tag{159}$$

it holds that

$$\forall h \in [H]: \quad V_h^{\pi,\sigma,w,\theta}(x_w) = x_h^{\pi,w,\theta} V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1 - x_h^{\pi,w,\theta}) V_{h+1}^{\pi,\sigma,w,\theta}(x_w), \tag{160a}$$

$$\forall (s,h) \in \mathcal{Y} \times [H]: \quad V_h^{\pi,\sigma,w,\theta}(s) = 1 + (1-\sigma) V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w). \tag{160b}$$

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

$$\begin{aligned}&\pi_h^{\star,w,\theta}(\theta_h \,|\, x_w) = \pi_h^{\star,w,\theta}(\theta_h \,|\, y_w) = 1,\\ &\pi_h^{\star,w,\theta}(\theta_h^{\mathsf{base}} \,|\, s) = \pi_h^{\star,w,\theta}(\theta_h^{\mathsf{base}} \,|\, s^{x\to y}) = 1, \quad \forall s \in \mathcal{X}\setminus\{x_w\}\end{aligned} \tag{161a}$$

and

$$V_h^{\star,\sigma,w,\theta}(x_w) = \underline{p} V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1-\underline{p}) V_{h+1}^{\pi,\sigma,w,\theta}(x_w). \tag{162}$$

D.2. Establishing the lower bound

Recall our goal: for any policy estimator $\widehat{\pi}$ computed based on the dataset with N samples, we plan to control the quantity

$$\max_{(w,\theta)\in\mathcal{W}\times\Theta^\star} \max_{s\in\mathcal{X}\cup\mathcal{Y}} \left\{ V_1^{\star,\sigma,w,\theta}(s) - V_1^{\widehat{\pi},\sigma,w,\theta}(s) \right\} \geq \max_{(w,\theta)\in\mathcal{W}\times\Theta^\star} \max_{s\in\mathcal{X}} \left\{ V_1^{\star,\sigma,w,\theta}(s) - V_1^{\widehat{\pi},\sigma,w,\theta}(s) \right\}. \tag{163}$$

Step 1: converting the goal to estimate (w,θ). Towards this, we make the following essential claim which shall be verified in Appendix D.3.2: letting

$$\varepsilon \leq \begin{cases} \frac{c_2}{H}, & \text{if } \sigma \leq \frac{c_2}{2H} \\ 1 & \text{otherwise} \end{cases} \tag{164}$$

and

$$\Delta = c_5 \begin{cases} \frac{\varepsilon}{H^2}, & \text{if } \sigma \leq \frac{c_2}{2H} \\ \frac{\sigma\varepsilon}{H} & \text{otherwise} \end{cases} \tag{165}$$

which satisfies (153), it leads to that for any policy π obeying

$$\sum_{h=1}^{H} \left\| \widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w) \right\|_1 \geq \frac{H}{8}, \tag{166}$$

one has

$$V_1^{\star,\sigma,w,\theta}(x_w) - V_1^{\widehat{\pi},\sigma,w,\theta}(x_w) > \varepsilon. \tag{167}$$

Now we are ready to convert the estimation of an optimal policy to estimate (w,θ). Towards this, we denote $\mathbb{P}_{w,\theta}$ as the probability distribution when the RMDP is $\mathcal{M}_w^\theta$ for any $(w,\theta) \in \mathcal{W}\times\Theta^\star$. In addition, we represent the subset of $\mathcal{M}(\mathcal{W},\Theta)$ excluding the ones associated with some $w \in \mathcal{W}$ as below:

$$\mathcal{G}_{-w} \coloneqq \mathcal{W}\setminus\{w\} \times \Theta^\star. \tag{168}$$

Then, for any $(w,\theta) \in \mathcal{W}\times\Theta^\star$, suppose there exists a policy $\widehat{\pi}$ that achieves

$$\mathbb{P}_{w,\theta}\left\{ V_1^{\star,\sigma,w,\theta}(x_w) - V_1^{\widehat{\pi},\sigma,w,\theta}(x_w) \leq \varepsilon \right\} \geq \frac{3}{4}, \tag{169}$$

which in view of (167) indicates that we necessarily have

$$\mathbb{P}_{w,\theta}\left\{ \sum_{h=1}^{H} \left\| \widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w) \right\|_1 < \frac{H}{8} \right\} \geq \frac{3}{4}. \tag{170}$$

Consequently, taking $\widetilde{\theta} = \arg\min_{\theta\in\Theta} \sum_{h=1}^{H} \left\| \widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w) \right\|_1$, we are motivated to construct the following estimate of (w,θ):

$$(\widehat{w},\widehat{\theta}) \begin{cases} = (w,\widetilde{\theta}) & \text{if} \quad \widetilde{\theta} \in \Theta^\star \\ \in \mathcal{G}_{-w} & \text{if} \quad \widetilde{\theta} = \Theta\setminus\Theta^\star = \theta^{\mathsf{base}}. \end{cases} \tag{171}$$

Then let us focus on the first kind of scenarios in (171) when $\widetilde{\theta} \in \Theta^\star$ so that we have the hope to estimate (w, θ) correctly. Namely, if $\sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w)\right\|_1 < \frac{H}{8}$ holds for some $\theta \in \Theta^\star$, then for any $\theta' \in \Theta^\star$ obeying $\theta' \neq \theta$, one has

$$\begin{aligned}\sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta'}(\cdot \,|\, x_w)\right\|_1 &\geq \sum_{h=1}^H \left\|\pi_h^{\star,w,\theta}(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta'}(\cdot \,|\, x_w)\right\|_1 - \sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w)\right\|_1 \\ &> \frac{H}{4} - \frac{H}{8} = \frac{H}{8}, \end{aligned} \tag{172}$$

where the first inequality holds by the triangle inequality, and the last inequality follows from the assumption $\sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w)\right\|_1 < \frac{H}{8}$ and the separation property of $\theta \in \Theta$ (see (145)). Similarly, It shows that we have $(\widehat{w}, \widehat{\theta}) = (w, \theta)$ if

$$\sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w)\right\|_1 < \frac{H}{8} < \sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta'}(\cdot \,|\, x_w)\right\|_1 \tag{173}$$

holds for all $(w', \theta') \in \mathcal{W} \times \Theta$ that $(w', \theta') \neq (w, \theta)$. It is clear that the above equation can be directly achieved when $\sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w)\right\|_1 < \frac{H}{8}$, which gives

$$\mathbb{P}_{w,\theta}\left[(\widehat{w}, \widehat{\theta}) = (w, \theta)\right] \geq \mathbb{P}_{w,\theta}\left\{\sum_{h=1}^H \left\|\widehat{\pi}_h(\cdot \,|\, x_w) - \pi_h^{\star,w,\theta}(\cdot \,|\, x_w)\right\|_1 < \frac{H}{8}\right\} \geq \frac{3}{4}. \tag{174}$$

Step 2: developing the probability of error in testing multiple hypotheses. Before proceeding, we discuss the data generation choices of the dataset $\mathcal{D}$. Recall that each RMDP inside the set $\mathcal{M}(\mathcal{W}, \Theta)$ under testing has two classes of states $\mathcal{X}$ and $\mathcal{Y}$, with each has $|\mathcal{W}| = SA$ components. Noticing that accordingly, $\mathcal{M}(\mathcal{W}, \Theta)$ consists of $|\mathcal{W}|$ subset, with each $\{\mathcal{M}_w^\theta\}_{\theta \in \Theta^\star}$ constructed symmetrically around one pair of state $(x_w, y_w) \in \mathcal{X} \times \mathcal{Y}$, respectively. Therefore, at each time step h, it is clear that the dataset are supposed to be generated uniformly by the transition kernels on each pair of states $(x_w, y_w) \in \mathcal{X} \times \mathcal{Y}$ to maximize the information gain. Namely, the dataset $\mathcal{D}$ has in total $\frac{N}{|\mathcal{W}|H} = \frac{N}{SAH}$ samples for the two states $(x_w, y_w) \in \mathcal{X} \times \mathcal{Y}$ at each time step $h \in [H]$.

Now we turn to the hypothesis testing problem over $(w, \theta) \in \mathcal{W} \times \Theta^\star$. We shall develop the information theoretical lower bound for the probability of error. In particular, we consider the minimax probability of error defined as follows:

$$p_{\mathrm{e}} \coloneqq \inf_{(\widehat{w}, \widehat{\theta})} \max_{(w,\theta) \in \mathcal{W} \times \Theta^\star} \left\{\mathbb{P}_{w,\theta}\big((\widehat{w}, \widehat{\theta}) \neq (w, \theta)\big)\right\}, \tag{175}$$

where the infimum is taken over all possible tests $(\widehat{w}, \widehat{\theta})$ constructed from the dataset.

To continue, armed with the dataset $\mathcal{D}$ with N samples generated independently, we denote $\mu^{w,\theta}$ (resp. $\mu_h^{w,\theta}(s, a)$) as the distribution vector (resp. distribution) of each sample tuple (s_h, a_h, s_h') at time step h under the nominal transition kernel $P^{w,\theta}$ associated with $\mathcal{M}_w^\theta$. With this in mind, combined with Fano's inequality from Tsybakov (2009, Theorem 2.2) and the additivity of the KL divergence (cf. Tsybakov (2009, Page 85)), we obtain

$$\begin{aligned} p_{\mathrm{e}} &\geq 1 - N \frac{\max\limits_{(w,\theta),(\widetilde{w},\widetilde{\theta}) \in \mathcal{W} \times \Theta^\star, (w,\theta) \neq (\widetilde{w},\widetilde{\theta})} \mathsf{KL}\big(\mu^{w,\theta} \,|\, \mu^{w,\theta}\big) + \log 2}{\log |\mathcal{W}||\Theta^\star|} \\ &\overset{\text{(i)}}{\geq} 1 - \frac{8N}{H} \max_{(w,\theta),(\widetilde{w},\widetilde{\theta}) \in \mathcal{W} \times \Theta^\star, (w,\theta) \neq (\widetilde{w},\widetilde{\theta})} \mathsf{KL}\big(\mu^{w,\theta} \,|\, \mu^{w,\theta}\big) - \frac{\log 2}{H} \\ &\overset{\text{(ii)}}{\geq} \frac{1}{2} - \frac{8N}{H} \max_{(w,\theta),(\widetilde{w},\widetilde{\theta}) \in \mathcal{W} \times \Theta^\star, (w,\theta) \neq (\widetilde{w},\widetilde{\theta})} \mathsf{KL}\big(\mu^{w,\theta} \,|\, \mu^{w,\theta}\big) \end{aligned} \tag{176}$$

where (i) and (ii) holds by $|\mathcal{W}||\Theta^\star| \geq 2(e^{H/8} - 1) \geq e^{H/8}$ as long as $H \geq 16 \log 2$.

To continue, applying the chain rule of the KL divergence (Duchi, 2018, Lemma 5.2.8) with the dataset $\mathcal{D}$ generated independently yields:

$$\begin{aligned}\mathsf{KL}\big(\mu^{w,\theta}\,|\,\mu^{w,\theta}\big) &= \sum_{h=1}^{H} \mathop{\mathbb{E}}_{(s,a)\sim\mu_h^{w,\theta}(s,a)} \Big[\mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,s,a)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,s,a)\big)\Big] \\ &\overset{\text{(i)}}{=} \sum_{h=1}^{H} \sum_{s\in\{x_w,x_{\widetilde{w}}\},a\in\{0,1\}} \mu_h^{w,\theta}(s,a)\Big[\mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,,s,a)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,,s,a)\big)\Big] \\ &\le \frac{1}{SAH}\sum_{h=1}^{H} \sum_{s\in\{x_w,x_{\widetilde{w}}\},a\in\{0,1\}} \Big[\mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,,s,a)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,,s,a)\big)\Big], \end{aligned} \tag{177}$$

where (i) follows from the fact $P_h^{w,\theta}(\cdot\,|\,,s,a)$ and $P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,,s,a)$ only differs from each other on state $x_w, x_{\widetilde{w}}$ (see the definitions in (147)), and the last inequality holds by noticing $\mu_h^{w,\theta}(s,a) \le \sum_{a\in\{0,1\}} \mu_h^{w,\theta}(s,a) = \frac{1}{SAH}$.

Consequently, now we turn to focus on terms in (177) in different cases of the uncertainty level σ.

- When $0 < \sigma \le \frac{c_2}{2H}$. When $w = \widetilde{w}$, it is clear that

$$\sum_{s\in\{x_w,x_{\widetilde{w}}\},a\in\{0,1\}} \mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,,s,a)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,,s,a)\big) = 0 \tag{178}$$

  as long as $\theta_h = \widetilde{\theta}_h$. Then if $\theta_h \neq \widetilde{\theta}_h$, without loss of generality, we suppose $\theta_h = 0$ and $\widetilde{\theta}_h = 1$, which indicates

$$P_h^{w,\theta}(0\,|\,x_w,0) = 1-p \quad \text{and} \quad P_h^{w,\widetilde{\theta}}(0\,|\,x_w,0) = 1-q. \tag{179}$$

  Applying Lemma D.1 gives

$$\begin{aligned}\mathsf{KL}\big(P_h^{w,\theta}(0\,|\,x_w,0)\;\|\;P_h^{w,\widetilde{\theta}}(0\,|\,x_w,0)\big) &\le \frac{(p-q)^2}{q(1-q)} \overset{\text{(i)}}{=} \frac{\Delta^2}{q(1-q)} \\ &\overset{\text{(ii)}}{=} \frac{(c_5)^2\varepsilon^2}{H^4 q(1-q)} \le \frac{4(c_5)^2\varepsilon^2}{c_2 H^3}, \end{aligned} \tag{180}$$

  where (i) and (ii) follows from the definitions in (149) or (165), and the last inequality arises from $q = p - \Delta \ge \frac{c_2}{2H}$ (see (153)) and $1-q \ge 1-p \ge 1-\frac{c_2}{H} \ge \frac{1}{2}$.

  The same bound can be established for $\mathsf{KL}\big(P_h^{w,\theta}(0\,|\,x_w,1)\;\|\;P_h^{w,\widetilde{\theta}}(0\,|\,x_w,1)\big)$. In addition, it is easily verified that when $w \neq \widetilde{w}$ and $\theta_h \neq \theta_h^{\mathsf{base}}$ (resp. $\widetilde{\theta}_h \neq \theta_h^{\mathsf{base}}$), the same bound can be developed for $\mathsf{KL}\big(P_h^{w,\theta}(0\,|\,x_w,0)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(0\,|\,x_w,0)\big)$ and $\mathsf{KL}\big(P_h^{w,\theta}(0\,|\,x_w,1)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(0\,|\,x_w,1)\big)$ (resp. $\mathsf{KL}\big(P_h^{w,\theta}(0\,|\,x_{\widetilde{w}},0)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(0\,|\,x_{\widetilde{w}},0)\big)$ and $\mathsf{KL}\big(P_h^{w,\theta}(0\,|\,x_{\widetilde{w}},1)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(0\,|\,x_{\widetilde{w}},1)\big)$).

  Summing up the results with the fact in (180), we arrive at

$$\sum_{s\in\{x_w,x_{\widetilde{w}}\},a\in\{0,1\}} \mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,,s,a)\;\|\;P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,,s,a)\big) \le \frac{16(c_5)^2\varepsilon^2}{c_2 H^3}. \tag{181}$$

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

$$\begin{aligned}\mathsf{KL}\big(P_h^{w,\theta}(0\,|\,x_w,0)\;\|\;P_h^{w,\widetilde{\theta}}(0\,|\,x_w,0)\big) &\le \frac{(p-q)^2}{q(1-q)} \overset{\text{(i)}}{=} \frac{\Delta^2}{q(1-q)} \\ &\overset{\text{(ii)}}{=} \frac{(c_5)^2\sigma^2\varepsilon^2}{H^2 q(1-q)} \le \frac{2(c_5)^2\sigma\varepsilon^2}{c_0 H^2}, \end{aligned} \tag{182}$$

where (i) and (ii) follows from the definitions in (149) or (165). Here, the last inequality arises from

$$1-q \geq 1-p = 1-(1+\frac{c_1}{H})\sigma \overset{\text{(i)}}{\geq} c_0 - \frac{c_1}{H} \overset{\text{(ii)}}{\geq} \frac{c_0}{2}$$
$$p \geq q = p - \Delta \overset{\text{(iii)}}{\geq} \sigma, \tag{183}$$

where (ii) holds by the definition of $c_1 = \frac{c_0}{2}$, and (iii) follows from (157). Consequently, we arrive at

$$\sum_{s\in\{x_w, x_{\widetilde{w}}\}, a\in\{0,1\}} \mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,s,a) \,\|\, P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,s,a)\big) \leq \frac{8(c_5)^2\sigma\varepsilon^2}{c_0 H^2}. \tag{184}$$

Summing up (181) and (184), we achieve for any $(w,\theta), (\widetilde{w},\widetilde{\theta}) \in \mathcal{W}\times\Theta^\star$ with $(w,\theta)\neq(\widetilde{w},\widetilde{\theta})$ and any time step $h\in[H]$

$$\sum_{s\in\{x_w, x_{\widetilde{w}}\}, a\in\{0,1\}} \mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,s,a) \,\|\, P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,s,a)\big) \leq \frac{16(c_5)^2\varepsilon^2}{c_0 c_2 H^2}\max\{\sigma, 1/H\}. \tag{185}$$

Plugging (185) back to (177) and then (176) leads to the following fact:

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{2} - \frac{8N}{H} \max_{(w,\theta),(\widetilde{w},\widetilde{\theta})\in\mathcal{W}\times\Theta^\star,(w,\theta)\neq(\widetilde{w},\widetilde{\theta})} \mathsf{KL}\big(\mu^{w,\theta}\,|\,\mu^{w,\theta}\big) \\ &\geq \frac{1}{2} - \frac{8N}{H} \max_{(w,\theta),(\widetilde{w},\widetilde{\theta})\in\mathcal{W}\times\Theta^\star,(w,\theta)\neq(\widetilde{w},\widetilde{\theta})} \frac{1}{SAH}\sum_{h=1}^{H}\sum_{s\in\{x_w,x_{\widetilde{w}}\},a\in\{0,1\}} \Big[\mathsf{KL}\big(P_h^{w,\theta}(\cdot\,|\,s,a) \,\|\, P_h^{\widetilde{w},\widetilde{\theta}}(\cdot\,|\,s,a)\big)\Big] \\ &\geq \frac{1}{2} - \frac{128N(c_5)^2\varepsilon^2}{c_0 c_2 SAH^3}\max\{\sigma, 1/H\} \geq \frac{1}{4} \end{aligned} \tag{186}$$

as long as the sample size N of the dataset is selected as

$$N \leq \frac{c_0 c_2 SAH^3 \min\{1/\sigma, H\}}{512(c_5)^2\varepsilon^2}. \tag{187}$$

Step 3: summing up the results together. We suppose that there exists an estimator $\widehat{\pi}$ such that

$$\max_{(w,\theta)\in\mathcal{W}\times\Theta^\star} \mathbb{P}_{w,\theta}\left[\max_{s\in\mathcal{X}\cup\mathcal{Y}}\left\{V_1^{\star,\sigma,w,\theta}(s) - V_1^{\widehat{\pi},\sigma,w,\theta}(s)\right\} \geq \varepsilon\right] < \frac{1}{4}, \tag{188}$$

then according to (163), we necessarily have

$$\forall w\in\mathcal{W}: \quad \max_{\theta\in\Theta^\star} \mathbb{P}_{w,\theta}\left[\left\{V_1^{\star,\sigma,w,\theta}(x_w) - V_1^{\widehat{\pi},\sigma,w,\theta}(x_w)\right\} \geq \varepsilon\right] < \frac{1}{4}. \tag{189}$$

To meet (189) for any $w\in\mathcal{W}$, we require

$$\forall\theta\in\Theta^\star: \mathbb{P}_{w,\theta}\left\{V_1^{\star,\sigma,w,\theta}(x_w) - V_1^{\widehat{\pi},\sigma,w,\theta}(x_w) < \varepsilon\right\} \geq \frac{3}{4}, \tag{190}$$

which in view of (167) indicates that we necessarily have

$$\forall\theta\in\Theta^\star: \quad \mathbb{P}_{w,\theta}\left\{\sum_{h=1}^{H}\big\|\widehat{\pi}_h(\cdot\,|\,x_w) - \pi_h^{\star,w,\theta}(\cdot\,|\,x_w)\big\|_1 < \frac{H}{8}\right\} \geq \frac{3}{4}. \tag{191}$$

As a consequence, (174) indicates

$$\forall\theta\in\Theta^\star: \mathbb{P}_{w,\theta}\Big[(\widehat{w},\widehat{\theta}) = (w,\theta)\Big] \geq \frac{3}{4}. \tag{192}$$

Applying the fact in (192) to all $w \in \mathcal{W}$ leads to one necessarily has

$$\forall (w,\theta) \in \mathcal{W} \times \Theta^\star : \quad \mathbb{P}_{w,\theta}\left[(\widehat{w},\widehat{\theta}) = (w,\theta)\right] \geq \frac{3}{4} \tag{193}$$

to achieve (188).

However, this would contract with (186) as long as the sample size condition in (187) is satisfied. Thus, if the sample size obeys the condition (187), we can't achieve an estimate $\widehat{\pi}$ that satisfies (188), which complete the proof.

D.3. Proof of the auxiliary facts

D.3.1. Proof of Lemma D.2

As all RMDPs within $\mathcal{M}(\mathcal{W},\Theta)$ are constructed analogously over each $w \in \mathcal{W}$ and $\theta \in \Theta^\star$, in this section, we shall focus on one specific RMDP $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W},\Theta)$, whose facts can be carried on for all other RMDPs in $\mathcal{M}(\mathcal{W},\Theta)$ directly.

Step 1: ordering the robust value function over different states. Before proceeding, we introduce several facts and notations that are useful throughout this section. First, we observe that for any $\mathcal{M}_w^\theta$ and any policy π: at the final step $H+1$,

$$\forall s \in \mathcal{X} \cup \mathcal{Y} : \quad V_{H+1}^{\pi,\sigma,w,\theta}(s) = 0. \tag{194}$$

Then for the step H, we can easily verified that

$$\begin{aligned}
\forall s \in \mathcal{Y} : \quad V_H^{\pi,\sigma,w,\theta}(s) &= \mathbb{E}_{a\sim\pi_H(\cdot\,|\,s)}\left[r_H(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{H,s,a}^{w,\theta})} \mathcal{P}V_{H+1}^{\pi,\sigma,w,\theta}\right] = 1 \\
\forall s \in \mathcal{X} : \quad V_H^{\pi,\sigma,w,\theta}(s) &= \mathbb{E}_{a\sim\pi_H(\cdot\,|\,s)}\left[r_H(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{H,s,a}^{w,\theta})} \mathcal{P}V_{H+1}^{\pi,\sigma,w,\theta}\right] = 0,
\end{aligned} \tag{195}$$

which holds by (194) and the definition of the reward function (see (150)). The above fact directly indicates that

$$\begin{aligned}
\forall (s,s') \in \mathcal{X}\setminus\{x_w\} \times \mathcal{Y} : \quad & \min_{\widetilde{s}\in\mathcal{S}} V_H^{\pi,\sigma,w,\theta}(\widetilde{s}) = V_H^{\pi,\sigma,w,\theta}(x_w) \leq V_H^{\pi,\sigma,w,\theta}(s) < V_H^{\pi,\sigma,w,\theta}(s'), \\
\forall (s,s') \in \mathcal{Y} \times \mathcal{Y} : \quad & V_H^{\pi,\sigma,w,\theta}(s) = V_H^{\pi,\sigma,w,\theta}(s').
\end{aligned} \tag{196}$$

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

$$\begin{aligned}
\forall (h,s,s') \in [H] \times \mathcal{X}\setminus\{x_w\} \times \mathcal{Y} : \quad & V_h^{\pi,\sigma,w,\theta}(x_w) \leq V_h^{\pi,\sigma,w,\theta}(s) < V_h^{\pi,\sigma,w,\theta}(s') \\
\forall (s,s') \in \mathcal{Y} \times \mathcal{Y} : \quad & V_h^{\pi,\sigma,w,\theta}(s) = V_h^{\pi,\sigma,w,\theta}(s').
\end{aligned} \tag{197}$$

Note that the base case when the time step is $H+1$ is verified in (196). Assuming that the following fact at time step $h+1$ holds

$$\begin{aligned}
\forall (s,s') \in \mathcal{X}\setminus\{x_w\} \times \mathcal{Y} : \quad & \min_{\widetilde{s}\in\mathcal{S}} V_{h+1}^{\pi,\sigma,w,\theta}(\widetilde{s}) = V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \leq V_{h+1}^{\pi,\sigma,w,\theta}(s) < V_{h+1}^{\pi,\sigma,w,\theta}(s'), \\
\forall (s,s') \in \mathcal{Y} \times \mathcal{Y} : \quad & V_{h+1}^{\pi,\sigma,w,\theta}(s) = V_{h+1}^{\pi,\sigma,w,\theta}(s'),
\end{aligned} \tag{198}$$

the rest of the proof focuses on proving the same property for time step h. For RMDP $\mathcal{M}_w^\theta \in \mathcal{M}(\mathcal{W},\Theta)$ and any policy π, we characterize the robust value function of different states separately:

- *For state $s \in \mathcal{Y}$.* We observe that for any $s \in \mathcal{Y}$,

$$\begin{aligned}
V_h^{\pi,\sigma,w,\theta}(s) &= \mathbb{E}_{a\sim\pi_h(\cdot\,|\,s)}\left[r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{h,s,a}^{w,\theta})} \mathcal{P}V_{h+1}^{\pi,\sigma,w,\theta}\right] \\
&\overset{\text{(i)}}{=} 1 + \mathbb{E}_{a\sim\pi_h(\cdot\,|\,s)}\left[\underline{P}_h^{w,\theta}(s\,|\,s,a)V_{h+1}^{\pi,\sigma,w,\theta}(s)\right] + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w)
\end{aligned}$$

$$= 1 + (1-\sigma)V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w), \tag{199}$$

where (i) holds by $r_h(s,a) = 1$ for all $s \in \mathcal{Y}$ (see (150)), the fact that $\min_{\widetilde{s}\in\mathcal{S}} V_{h+1}^{\pi,\sigma,w,\theta}(\widetilde{s}) = V_{h+1}^{\pi,\sigma,w,\theta}(x_w)$ induced by the induction assumption (cf. (198)) and the definition of $\underline{P}_h^{w,\theta}(s \,|\, s,a)$ in (155), and the last equality follows from $P^{w,\theta}(s \,|\, s,a) = 1$ for all $(s,a) \in \mathcal{Y} \times \mathcal{A}_{\mathcal{M}}$. Resorting to the induction assumption in (198), we have

$$\forall (s,s') \in \mathcal{Y} \times \mathcal{Y}: \quad V_h^{\pi,\sigma,w,\theta}(s) = V_h^{\pi,\sigma,w,\theta}(s'). \tag{200}$$

- *For state x_w.* First, the robust value function at state x_w obeys

$$\begin{aligned}
&V_h^{\pi,\sigma,w,\theta}(x_w) \\
&= \mathbb{E}_{a\sim\pi_h(\cdot\,|\,x_w)}\left[r_h(x_w,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{w,\theta}_{h,x_w,a})} \mathcal{P}V_{h+1}^{\pi,\sigma,w,\theta}\right] \\
&\overset{\text{(i)}}{=} 0 + \pi_h(\theta_h \,|\, x_w) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{w,\theta}_{h,x_w,\theta_h})} \mathcal{P}V_{h+1}^{\pi,\sigma,w,\theta} + \pi_h(1-\theta_h \,|\, x_w) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^{w,\theta}_{h,x_w,1-\theta_h})} \mathcal{P}V_{h+1}^{\pi,\sigma,w,\theta} \\
&\overset{\text{(ii)}}{=} \pi_h(\theta_h \,|\, x_w)\Big[\underline{p}V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + \big(1-\underline{p}\big)\, V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\Big] \\
&\quad + \pi_h(1-\theta_h \,|\, x_w)\Big[\underline{q}V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + \big(1-\underline{q}\big)\, V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\Big] \\
&\overset{\text{(iii)}}{=} x_h^{\pi,w,\theta}V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1-x_h^{\pi,w,\theta})V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \qquad (201)\\
&\le (1-\sigma)V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w). \qquad (202)
\end{aligned}$$

where (i) uses the definition of the robust value function and the reward function in (150), (ii) uses the induction assumption in (198) so that the minimum is attained by picking the choice specified in (156) to absorb probability mass to state x_w, and (iii) holds by plugging in the definition (159) of $x_h^{\pi,w,\theta}$ in (iii). Finally, the last inequality follows from the fact that function $f(x) \coloneqq xV_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1-x)V_{h+1}^{\pi,\sigma,w,\theta}(x_w)$ is monotonically increasing with x since $V_{h+1}^{\pi,\sigma,w,\theta}(y_w) > V_{h+1}^{\pi,\sigma,w,\theta}(x_w)$ (see the induction assumption (198)), and the fact $x_h^{\pi,w,\theta} \le 1-\sigma$.

- *For state $s \in \mathcal{X}\setminus\{x_w\}$.* Then we consider other states $s \in \mathcal{X}\setminus\{x_w\}$. Before proceeding, analogous to (159), we define

$$x_{\mathsf{base}}^s = (\underline{p}+\Delta)\pi_h(\theta_h^{\mathsf{base}} \,|\, s) + (\underline{q}+\Delta)\pi_h(1-\theta_h^{\mathsf{base}} \,|\, s). \tag{203}$$

Recall that the nominal transition kernel at any state $s \in \mathcal{X}\setminus\{x_w\}$ are the same $\{P^\star_{h,s,a}\}_{h\in[H]}$ for all $a \in \mathcal{A}_{\mathcal{W}}$ associated with the basic $\theta^{\mathsf{base}} \in \Theta$ (see the definitions of the transition kernels in (147) and (148)). Consequently, for any $s \in \mathcal{X}\setminus\{x_w\}$, following the same argument pipeline of (202), we arrive at

$$\begin{aligned}
V_h^{\pi,\sigma,w,\theta}(s) &= \pi_h(\theta_h^{\mathsf{base}} \,|\, s)\Big[(\underline{p}+\Delta)V_{h+1}^{\pi,\sigma,w,\theta}(s^{x\to y}) + (1-p-\Delta)\, V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\Big] \\
&\quad + \pi_h(1-\theta_h^{\mathsf{base}} \,|\, s)\Big[(\underline{q}+\Delta)V_{h+1}^{\pi,\sigma,w,\theta}(s^{x\to y}) + (1-p)\, V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\Big] \\
&= x_{\mathsf{base}}^s V_{h+1}^{\pi,\sigma,w,\theta}(s^{x\to y}) + (1-x_{\mathsf{base}}^s-\sigma)V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \qquad (204)\\
&\overset{\text{(i)}}{=} x_{\mathsf{base}}^s V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1-x_{\mathsf{base}}^s-\sigma)V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \qquad (205)\\
&< (1-\sigma)V_{h+1}^{\pi,\sigma,w,\theta}(s^{x\to y}) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(s), \qquad (206)
\end{aligned}$$

where (i) holds by $V_{h+1}^{\pi,\sigma,w,\theta}(s) = V_{h+1}^{\pi,\sigma,w,\theta}(s')$ for any two states $s, s' \in \mathcal{Y}$ (see (202)), and the last inequality holds by $V_{h+1}^{\pi,\sigma,w,\theta}(s) < V_{h+1}^{\pi,\sigma,w,\theta}(s^{x\to y})$ induced by the induction assumption in (198).

In addition, to compare the robust value function $V_h^{\pi,\sigma,w,\theta}(x_w)$ to that of other states $s \in \mathcal{X}\setminus\{x_w\}$, we recall the definitions in (159) and then introduce the following fact

$$\begin{aligned}
x_h^{\pi,w,\theta} &= \underline{p}\pi_h(\theta_h \,|\, x_w) + \underline{q}\pi_h(1-\theta_h \,|\, x_w) \\
&\le \underline{p} \le (\underline{p}+\Delta)\pi_h(\theta_h^{\mathsf{base}} \,|\, s) + \underline{p}\pi_h(1-\theta_h^{\mathsf{base}} \,|\, s)
\end{aligned}$$

$$= (\underline{p} + \Delta)\pi_h(\theta_h^{\mathsf{base}} \,|\, s) + (\underline{q} + \Delta)\pi_h(1 - \theta_h^{\mathsf{base}} \,|\, s) = x_{\mathsf{base}}^s, \tag{207}$$

which comes from the fact $p \geq q$ and the facts in (156) and (157).

With this in mind, continuing from (201), we arrive at that for any $s \in \mathcal{X}$:

$$\begin{aligned} V_h^{\pi,\sigma,w,\theta}(x_w) &= x_h^{\pi,w,\theta} V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1 - x_h^{\pi,w,\theta}) V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \\ &\leq x_{\mathsf{base}}^s V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1 - x_{\mathsf{base}}^s) V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \\ &\leq x_{\mathsf{base}}^s V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1 - x_{\mathsf{base}}^s - \sigma) V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \\ &= V_h^{\pi,\sigma,w,\theta}(s) \end{aligned} \tag{208}$$

where the last equality holds by (205).

Summing up (208), then (199), and (206), we verify the induction property at time step h as below

$$\forall (s, s') \in \mathcal{X} \setminus \{x_w\} \times \mathcal{Y} : V_h^{\pi,\sigma,w,\theta}(x_w) \leq V_h^{\pi,\sigma,w,\theta}(s) < V_h^{\pi,\sigma,w,\theta}(s'). \tag{209}$$

Combined above results with (200), we confirm the claim in (197).

Step 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 $\mathcal{X}$.* Recall (201)

$$V_h^{\pi,\sigma,w,\theta}(x_w) = x_h^{\pi,w,\theta} V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1 - x_h^{\pi,w,\theta}) V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \tag{210}$$

and the fact $V_{h+1}^{\pi,\sigma,w,\theta}(y_w) > V_{h+1}^{\pi,\sigma,w,\theta}(x_w)$ in (197). We observe that (210) is monotonicity increasing with respect to $x_h^{\pi,w,\theta}$, and $x_h^{\pi,w,\theta}$ is also increasing in $\pi_h(\theta_h \,|\, x_w)$ (refer to the fact $\underline{p} \geq \underline{q}$ since $p \geq q$; see (149) and (156)). Consequently, the optimal policy and optimal robust value function in state x_w thus obey

$$\begin{aligned} \forall h \in [H] : \quad & \pi_h^{\star,w,\theta}(\theta_h \,|\, x_w) = 1 \\ & V_h^{\star,\sigma,w,\theta}(x_w) = \underline{p} V_{h+1}^{\star,\sigma,w,\theta}(y_w) + \left[1 - \underline{p}\right] V_{h+1}^{\star,\sigma,w,\theta}(x_w). \end{aligned} \tag{211}$$

Similarly, for any state $s \in \mathcal{X} \setminus \{x_w\}$, recalling (205) yields

$$V_h^{\pi,\sigma,w,\theta}(s) = x_{\mathsf{base}}^s V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + (1 - x_{\mathsf{base}}^s - \sigma) V_{h+1}^{\pi,\sigma,w,\theta}(s) + \sigma V_{h+1}^{\pi,\sigma,w,\theta}(x_w), \tag{212}$$

which indicates $V_h^{\pi,\sigma,w,\theta}(s)$ achieves the maximum when $x_{\mathsf{base}}^s = (\underline{p} + \Delta)\pi_h(\theta_h^{\mathsf{base}} \,|\, s) + (\underline{q} + \Delta)\pi_h(1 - \theta_h^{\mathsf{base}} \,|\, s)$ attain the maximum. Therefore, the optimal policy in state s satisfies

$$\pi_h^{\star,w,\theta}(\theta_h^{\mathsf{base}} \,|\, s) = 1. \tag{213}$$

- *For states $s \in \mathcal{Y}$.* Recall the transitions in (147) and (148). Considering that the action does not influence the state transition for all states $s \in \mathcal{Y}$, without loss of generality, we choose the robust optimal policy obeying

$$\forall s \in \mathcal{Y} : \quad \pi_h^{\star,w,\theta}(\theta_h \,|\, s) = 1. \tag{214}$$

D.3.2. Proof of claim (167)

Recalling (160a) and (162), we first consider a more general form

$$\begin{aligned} & V_h^{\star,\sigma,w,\theta}(x_w) - V_h^{\pi,\sigma,w,\theta}(x_w) \\ &= \underline{p} V_{h+1}^{\star,\sigma,w,\theta}(y_w) + (1 - \underline{p}) V_{h+1}^{\star,\sigma,w,\theta}(x_w) - \left(x_h^{\pi,w,\theta} V_{h+1}^{\pi,\sigma,w,\theta}(y_w) + \left[1 - x_h^{\pi,w,\theta}\right] V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \right) \end{aligned}$$

$$
\begin{aligned}
&= \left(\underline{p} - x_h^{\pi,w,\theta}\right) V_{h+1}^{\star,\sigma,w,\theta}(y_w) + x_h^{\pi,w,\theta}\left(V_{h+1}^{\star,\sigma,w,\theta}(y_w) - V_{h+1}^{\pi,\sigma,w,\theta}(y_w)\right) \\
&\quad + (1-\underline{p})\left(V_{h+1}^{\star,\sigma,w,\theta}(x_w) - V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\right) - \left(\underline{p} - x_h^{\pi,w,\theta}\right) V_{h+1}^{\pi,\sigma,w,\theta}(x_w) \\
&= x_h^{\pi,w,\theta}\left(V_{h+1}^{\star,\sigma,w,\theta}(y_w) - V_{h+1}^{\pi,\sigma,w,\theta}(y_w)\right) + (1-\underline{p})\left(V_{h+1}^{\star,\sigma,w,\theta}(x_w) - V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\right) \\
&\quad + \left(\underline{p} - x_h^{\pi,w,\theta}\right)\left(V_{h+1}^{\star,\sigma,w,\theta}(y_w) - V_{h+1}^{\star,\sigma,w,\theta}(x_w)\right) \\
&\geq (1-\underline{p})\left(V_{h+1}^{\star,\sigma,w,\theta}(x_w) - V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\right) + \left(\underline{p} - x_h^{\pi,w,\theta}\right)\left(V_{h+1}^{\star,\sigma,w,\theta}(y_w) - V_{h+1}^{\star,\sigma,w,\theta}(x_w)\right) \\
&\geq (1-\underline{p})\left(V_{h+1}^{\star,\sigma,w,\theta}(x_w) - V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\right) \\
&\quad + \frac{1}{2}(p-q)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w) - \pi_h(\cdot\,|\,x_w)\right\|_1 \left(V_{h+1}^{\star,\sigma,w,\theta}(y_w) - V_{h+1}^{\star,\sigma,w,\theta}(x_w)\right)
\end{aligned} \tag{215}
$$

where the last inequality holds by applying (156) and deriving as follows:

$$
\begin{aligned}
\underline{p} - x_h^{\pi,w,\theta} &= \big(\underline{p} - \underline{q}\big)\big(1 - \pi_h(\theta_h\,|\,x_w)\big) = (p-q)\big(1-\pi_h(\theta_h\,|\,x_w)\big) \\
&= \frac{1}{2}(p-q)\big(1-\pi_h(\theta_h\,|\,x_w) + \pi_h(1-\theta_h\,|\,x_w)\big) = \frac{1}{2}(p-q)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w) - \pi_h(\cdot\,|\,x_w)\right\|_1.
\end{aligned} \tag{216}
$$

To further control (215), applying Lemma D.2 yields

$$
\begin{aligned}
&V_h^{\star,\sigma,w,\theta}(y_w) - V_h^{\star,\sigma,w,\theta}(x_w) \\
&= 1 + (1-\sigma)V_{h+1}^{\star,\sigma,w,\theta}(y_w) + \sigma V_{h+1}^{\star,\sigma,w,\theta}(x_w) - \left(\underline{p}V_{h+1}^{\star,\sigma,w,\theta}(y_w) + (1-\underline{p})V_{h+1}^{\star,\sigma,w,\theta}(x_w)\right) \\
&= 1 + (1-\underline{p}-\sigma)\left(V_{h+1}^{\star,\sigma,w,\theta}(y_w) - V_{h+1}^{\star,\sigma,w,\theta}(x_w)\right) \\
&= 1 + (1-p)\left(V_{h+1}^{\star,\sigma,w,\theta}(y_w) - V_{h+1}^{\star,\sigma,w,\theta}(x_w)\right) \\
&= \cdots = \sum_{j=0}^{H-h}(1-p)^j,
\end{aligned} \tag{217}
$$

where the penultimate equality holds by (156). Then, we consider two cases with respect to the uncertainty level σ to control (217), respectively:

- *When* $0 < \sigma \leq \frac{c_2}{2H}$. Recall $p = \begin{cases} \frac{c_2}{H}, & \text{if } \sigma \leq \frac{c_2}{2H} \\ 1 + \frac{c_1}{H}\sigma & \text{otherwise} \end{cases}$. In this case, applying (217), we have

$$
\begin{aligned}
&V_h^{\star,\sigma,w,\theta}(y_w) - V_h^{\star,\sigma,w,\theta}(x_w) \\
&= \sum_{j=0}^{H-h}(1-p)^j \geq \sum_{j=0}^{H-h}\left(1-\frac{c_2}{H}\right)^j = \frac{1-\left(1-\frac{c_2}{H}\right)^{H-h+1}}{c_2/H} \geq \frac{2c_2(H-h+1)}{3}
\end{aligned} \tag{218}
$$

  Here, the final inequality holds by observing

$$
\left(1-\frac{c_2}{H}\right)^{H-h+1} \leq \exp\left(-\frac{c_2(H-h+1)}{H}\right) \leq 1 - \frac{2c_2(H-h+1)}{3H} \tag{219}
$$

  where the first inequality holds by noticing $c_2 < 0.5$ and then $1 - x \leq \exp(-x)$, and the last inequality holds by $\exp(-x) \leq 1 - \frac{2x}{3}$ for any $0 \leq x \leq 1/2$.

  Plugging above fact in (218) back to (215), we arrive at

$$
\begin{aligned}
&V_h^{\star,\sigma,w,\theta}(x_w) - V_h^{\pi,\sigma,w,\theta}(x_w) \\
&\geq (1-\underline{p})\left(V_{h+1}^{\star,\sigma,w,\theta}(x_w) - V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\right)
\end{aligned}
$$

$$+\frac{1}{2}(p-q)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w)-\pi_h(\cdot\,|\,x_w)\right\|_1\frac{2c_2(H-h+1)}{3}. \tag{220}$$

Then invoking the assumption

$$\sum_{h=1}^{H}\left\|\pi_h(\cdot\,|\,x_w)-\pi_h^{\star,w,\theta}(\cdot\,|\,x_w)\right\|_1\geq\frac{H}{8} \tag{221}$$

in (166) and applying (220) recursively for $h=1,2,\cdots,H$ yields

$$\begin{aligned}
V_1^{\star,\sigma,w,\theta}(x_w)-V_1^{\pi,\sigma,w,\theta}(x_w)&\geq\frac{c_2}{3}\sum_{h=1}^{H}(1-\underline{p})^{h-1}(p-q)(H-h+1)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w)-\pi_h(\cdot\,|\,x_w)\right\|_1\\
&\overset{\text{(i)}}{\geq}\frac{c_2}{3}\sum_{h=1}^{H}(1-\frac{c_2}{H})^{h-1}(p-q)(H-h+1)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w)-\pi_h(\cdot\,|\,x_w)\right\|_1\\
&\overset{\text{(ii)}}{\geq}\frac{c_2}{6}\sum_{h=1}^{H}(p-q)(H-h+1)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w)-\pi_h(\cdot\,|\,x_w)\right\|_1\\
&\overset{\text{(iii)}}{=}\frac{c_2\Delta}{6}\sum_{h=1}^{H}h\left\|\pi_{H-h+1}^{\star,w,\theta}(\cdot\,|\,x_w)-\pi_{H-h+1}(\cdot\,|\,x_w)\right\|_1\\
&\overset{\text{(iv)}}{\geq}\frac{c_2\Delta}{6}\sum_{h=1}^{\lfloor H/16\rfloor}2h\geq\frac{c_2\Delta}{6}\lfloor H/16\rfloor\left(\lfloor H/16\rfloor+1\right),
\end{aligned} \tag{222}$$

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

$$\forall h\in[H]:\quad(1-\frac{c_2}{H})^{h-1}\geq(1-\frac{c_2}{H})^{H}\geq\frac{1}{2} \tag{223}$$

as long as $c_2\leq\frac{1}{2}$. Here, (iii) arises from the definition of p,q in (149); (iv) can be verified by the fact that for any series $0\leq x_1,x_2,\cdots,x_H\leq x_{\max}$ that obeys $\sum_{h=1}^{H}x_h\geq y$, one has

$$\sum_{h=1}^{H}x_h h\geq\sum_{h=1}^{\lfloor x_{\max}/y\rfloor}x_{\max}h, \tag{224}$$

and taking $x_h=\left\|\pi_{H-h+1}(\cdot\,|\,x_w)-\pi_{H-h+1}^{\star,w,\theta}(\cdot\,|\,x_w)\right\|_1\leq 2=x_{\max}$ and $y=\frac{H}{8}$.
Consequently, observed from (222), we have

$$V_1^{\star,\sigma,w,\theta}(x_w)-V_1^{\pi,\sigma,w,\theta}(x_w)\geq\frac{c_2\Delta}{6}\lfloor H/16\rfloor\left(\lfloor H/16\rfloor+1\right)\geq c_3\Delta H^2>\varepsilon \tag{225}$$

holds for some small enough constant c_3 and letting $\Delta=\frac{\varepsilon}{c_3H^2}$.

- *When* $\frac{c_2}{2H}<\sigma\leq 1-c_0$. Similarly, recalling $p=\begin{cases}\frac{c_2}{H}, & \text{if } \sigma\leq\frac{c_2}{2H}\\ \left(1+\frac{c_1}{H}\right)\sigma & \text{otherwise}\end{cases}$ and invoking (217) gives

$$\begin{aligned}
&V_h^{\star,\sigma,w,\theta}(y_w)-V_h^{\star,\sigma,w,\theta}(x_w)\\
&=\sum_{j=0}^{H-h}(1-p)^j=\sum_{j=0}^{H-h}\left(1-\left(1+\frac{c_1}{H}\right)\sigma\right)^j\\
&\geq\frac{1-\left(1-(1+\frac{c_1}{H})\sigma\right)^{H-h+1}}{(1+\frac{c_1}{H})\sigma}\geq\frac{c_2(H-h+1)}{3\sigma H},
\end{aligned} \tag{226}$$

where the final inequality holds by observing

$$\begin{aligned}\left(1-\left(1+\frac{c_1}{H}\right)\sigma\right)^{H-h+1} &\leq \exp\left(-\left(1+\frac{c_1}{H}\right)\sigma(H-h+1)\right)\\ &\overset{\text{(i)}}{\leq} \exp\left(-\frac{c_2}{2H}\left(1+\frac{c_1}{H}\right)(H-h+1)\right) \leq 1-\left(1+\frac{c_1}{H}\right)\frac{c_2(H-h+1)}{3H}. \end{aligned} \tag{227}$$

Here, (i) holds by observing $\frac{c_2}{2H} < \sigma$, and the last inequality holds by $\left(1+\frac{c_1}{H}\right) \leq 2$, $c_2 \leq 0.5$, and the fact $\exp(-x) \leq 1-\frac{2x}{3}$ for any $0 \leq x \leq 1/2$.

Plugging above fact in (226) back to (215) gives

$$\begin{aligned}&V_h^{\star,\sigma,w,\theta}(x_w) - V_h^{\pi,\sigma,w,\theta}(x_w)\\ &\geq (1-\underline{p})\left(V_{h+1}^{\star,\sigma,w,\theta}(x_w) - V_{h+1}^{\pi,\sigma,w,\theta}(x_w)\right)\\ &\quad + \frac{1}{2}(p-q)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w) - \pi_h(\cdot\,|\,x_w)\right\|_1 \frac{c_2(H-h+1)}{3\sigma H}. \end{aligned} \tag{228}$$

Following the same routine to achieve (222), applying (228) recursively for $h = 1, 2, \cdots, H$ gives

$$\begin{aligned}V_1^{\star,\sigma,w,\theta}(x_w) - V_1^{\pi,\sigma,w,\theta}(x_w) &\geq \sum_{h=1}^{H}(1-\underline{p})^{h-1}(p-q)\frac{c_2(H-h+1)}{6\sigma H}\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w) - \pi_h(\cdot\,|\,x_w)\right\|_1\\ &\overset{\text{(i)}}{=} \frac{c_2(p-q)}{6\sigma H}\sum_{h=1}^{H}(1-\frac{c_1}{H})^{h-1}(H-h+1)\left\|\pi_h^{\star,w,\theta}(\cdot\,|\,x_w) - \pi_h(\cdot\,|\,x_w)\right\|_1\\ &\overset{\text{(ii)}}{\geq} \frac{c_2\Delta}{12\sigma H}\lfloor H/16\rfloor\left(\lfloor H/16\rfloor+1\right) \end{aligned} \tag{229}$$

where (i) follows from $1-\underline{p} = 1-(p-\sigma) = 1-\frac{c_1}{H}\sigma$, and (ii) holds by letting $c_1 \leq \frac{1}{2}$ and following the same routine of (222).

Consequently, (229) yields

$$V_1^{\star,\sigma,w,\theta}(x_w) - V_1^{\pi,\sigma,w,\theta}(x_w) \geq \frac{c_2\Delta}{12\sigma H}\lfloor H/16\rfloor\left(\lfloor H/16\rfloor+1\right) \geq \frac{c_4\Delta H}{\sigma} > \varepsilon \tag{230}$$

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