Model-Based Reinforcement Learning Is Minimax-Optimal for Offline Zero-Sum Markov Games

Yuling Yan* Gen Li† Yuxin Chen† Jianqing Fan*

June 8, 2022

Abstract

This paper makes progress towards learning Nash equilibria in two-player zero-sum Markov games from offline data. Specifically, consider a γ-discounted infinite-horizon Markov game with S states, where the max-player has A actions and the min-player has B actions. We propose a pessimistic model-based algorithm with Bernstein-style lower confidence bounds — called VI-LCB-Game — that provably finds an ε-approximate Nash equilibrium with a sample complexity no larger than $\frac{C_{\mathsf{clipped}}^{\star} S(A+B)}{(1-\gamma)^3\varepsilon^2}$ (up to some log factor). Here, $C_{\mathsf{clipped}}^{\star}$ is some unilateral clipped concentrability coefficient that reflects the coverage and distribution shift of the available data (vis-à-vis the target data), and the target accuracy ε can be any value within $\big(0, \frac{1}{1-\gamma}\big]$. Our sample complexity bound strengthens prior art by a factor of $\min\{A, B\}$, achieving minimax optimality for the entire ε-range. An appealing feature of our result lies in algorithmic simplicity, which reveals the unnecessity of variance reduction and sample splitting in achieving sample optimality.

Keywords: zero-sum Markov games, Nash equilibrium, offline RL, model-based approach, unilateral coverage, curse of multiple agents, minimax optimality

Contents

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: `{yulingy,jqfan}@princeton.edu`.

†Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA; Email: `{ligen,yuxinc}@wharton.upenn.edu`.

1 Introduction

Multi-agent reinforcement learning (MARL), a subfield of reinforcement learning (RL) that involves multiple individuals interacting/competing with each other in a shared environment, has garnered widespread recent interest, partly sparked by its capability of achieving superhuman performance in game playing and autonomous driving (Baker et al., 2019; Berner et al., 2019; Brown and Sandholm, 2019; Jaderberg et al., 2019; Shalev-Shwartz et al., 2016; Vinyals et al., 2019). The coexistence of multiple players — whose own welfare might come at the expense of other parties involved — makes MARL inherently more intricate than the single-agent counterpart.

A standard framework to describe the environment and dynamics in competitive MARL is Markov games (MGs), which is generally attributed to Shapley (1953) (originally referred to as stochastic games). Given the conflicting needs of the players, a standard goal in Markov games is to seek some sort of steady-state solutions, with the Nash equilibrium (NE) being arguably the most prominent one. While computational intractability has been observed when calculating NEs in general-sum MGs and/or MGs with more than two players (Daskalakis, 2013; Daskalakis et al., 2009), an assortment of tractable algorithms have been put forward to solve two-player zero-sum Markov games. On this front, a large strand of recent works revolves around developing sample- and computation-efficient paradigms (Bai et al., 2020; Liu et al., 2021; Tian et al., 2021; Xie et al., 2020; Zhang et al., 2020a). What is particularly noteworthy here is the recent progress in overcoming the so-called "curse of multiple agents" (Bai et al., 2020; Jin et al., 2021a; Li et al., 2022a); that is, although the total number of joint actions exhibits exponential scaling in the number of agents, learnability of Nash equilibria becomes plausible even when the sample size scales only linearly with the maximum cardinality of the individual action spaces. See also Daskalakis et al. (2022); Jin et al. (2021a); Mao and Başar (2022); Song et al. (2021) for similar accomplishments in learning coarse correlated equilibria in multi-player general-sum MGs.

The aforementioned works permit online data collection either via active exploration of the environment, or through sampling access to a simulator. Nevertheless, the fact that real-time data acquisition might be unaffordable or unavailable — e.g., it could be time-consuming, costly, and/or unsafe in healthcare and autonomous driving — constitutes a major hurdle for widespread adoption of these online algorithms. This practical consideration inspires a recent flurry of studies collectively referred to as *offline RL* or *batch RL* (Kumar et al., 2020; Levine et al., 2020), with the aim of learning based on a historical dataset of logged interactions.

Data coverage for offline Markov games. The feasibility and efficiency of offline RL are largely governed by the coverage of the offline data in hand. On the one hand, if the available dataset covers all state-action pairs adequately, then there is sufficient information to guarantee learnability; on the other hand, full data coverage imposes an overly stringent requirement that is rarely fulfilled in practice, and is oftentimes wasteful in terms of data efficiency. Consequently, a recurring theme in offline RL gravitates around the quest for algorithms that work under minimal data coverage. Encouragingly, the recent advancement on this frontier (e.g., Rashidinejad et al. (2021); Xie et al. (2021)) uncovers the sufficiency of "single-policy" data coverage in single-agent RL; namely, offline RL becomes information theoretically feasible as soon as the historical data covers the part of the state-action space reachable by a single target policy.

Unfortunately, single-policy coverage is provably insufficient when it comes to Markov games, with negative evidence observed in Cui and Du (2022b). Instead, a sort of unilateral coverage — i.e., a condition that requires the data to cover not only the target policy pair but also any unilateral deviation from it — seems necessary to ensure learnability of Nash equilibria in two-player zero-sum MGs. Employing the so-called "unilateral concentrability coefficient" $C^\star$ to quantify such unilateral coverage as well as the degree of distribution shift (which we shall define shortly in Assumption 1), Cui and Du (2022b) demonstrated how to find ε-Nash solutions in a finite-horizon two-player zero-sum MG once the number of sample rollouts exceeds

$$\widetilde{O}\left(\frac{C^\star H^3 SAB}{\varepsilon^2}\right). \tag{1.1}$$

Here, S is the number of shared states, A and B represent respectively the number of actions of the max-player and the min-player, H stands for the horizon length, and the notation $\widetilde{O}(\cdot)$ denotes the orderwise scaling with all logarithmic dependency hidden.

Despite being an intriguing polynomial sample complexity bound, a shortfall of (1.1) lies in its unfavorable scaling with AB (i.e., the total number of joint actions), which is substantially larger than the total number of individual actions $A + B$. Whether it is possible to alleviate this curse of multiple agents — and if so, how to accomplish it — is the key question to be investigated in the current paper.

An overview of main results. The objective of this paper is to design a sample-efficient offline RL algorithm for learning Nash equilibria in two-player zero-sum Markov games, ideally breaking the curse of multiple agents. Focusing on γ-discounted infinite-horizon MGs, we propose a model-based paradigm — called VI-LCB-Game — that is capable of learning an ε-approximate Nash equilibrium with sample complexity

$$\widetilde{O}\left(\frac{C^\star_{\mathsf{clipped}} S\,(A+B)}{(1-\gamma)^3\,\varepsilon^2}\right),$$

where $C^\star_{\mathsf{clipped}}$ is the so-called "clipped unilateral concentrability coefficient" (to be formalized in Assumption 2) and always satisfies $C^\star_{\mathsf{clipped}} \le C^\star$. Our result strengthens prior theory in Cui and Du (2022b) by a factor of $\min\{A, B\}$ (if we view the horizon length H in finite-horizon MGs and the effective horizon $\frac{1}{1-\gamma}$ in the infinite-horizon counterpart as equivalent). To demonstrate that this bound is essentially un-improvable, we develop a matching minimax lower bound (up to some logarithmic factor), thus settling this problem. Our algorithm is a pessimistic variant of value iteration with carefully designed Bernstein-style penalties, which requires neither sample splitting nor sophisticated schemes like reference-advantage decomposition. The fact that our sample complexity result holds for the full ε-range (i.e., any $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$) unveils that sample efficiency is achieved without incurring any burn-in cost.

Finally, while finalizing the current paper, we became aware of an independent study Cui and Du (2022a) (posted to arXiv on June 1, 2022) that also manages to overcome the curse of multiple agents, which we shall elaborate on towards the end of Section 3.

Notation. Before proceeding, let us introduce several notation that will be used throughout. With slight abuse of notation, we shall use P to denote a probability transition kernel and the associated probability transition matrix exchangeably. We also use the notation μ exchangeably for a probability distribution and its associated probability vector (and we often do not specify whether μ is a row vector or column vector as long as it is clear from the context). For any two vectors $x = [x_i]_{i=1}^n$ and $y = [y_i]_{i=1}^n$, we use $x \circ y = [x_i y_i]_{i=1}^n$

to denote their Hadamard product, and we also define $x^2 = [x_i^2]_{i=1}^n$ in an entrywise fashion. For a finite set $\mathcal{S} = \{1, \cdots, S\}$, we let $\Delta(\mathcal{S}) \coloneqq \{x \in \mathbb{R}^S \mid 1^\top x = 1, x \geq 0\}$ represent the probability simplex over the set $\mathcal{S}$.

2 Problem formulation

In this section, we introduce the background of zero-sum Markov games, followed by a description of the offline dataset.

2.1 Preliminaries

Zero-sum two-player Markov games. Consider a discounted infinite-horizon zero-sum Markov Game (MG) (Littman, 1994; Shapley, 1953), as represented by the tuple $\mathcal{MG} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, \gamma)$. Here, $\mathcal{S} = \{1, \dots, S\}$ is the shared state space; $\mathcal{A} = \{1, \dots, A\}$ (resp. $\mathcal{B} = \{1, \dots, B\}$) is the action space of the max-player (resp. min-player); $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ is the (*a priori* unknown) probability transition kernel, where $P(s' \mid s, a, b)$ denotes the probability of transiting from state s to state s' if the max-player executes action a and the min-player chooses action b; $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ is the reward function, such that $r(s, a, b)$ indicates the immediate reward observed by both players in state s when the max-player takes action a and the min-player takes action b; and $\gamma \in (0, 1)$ is the discount factor, with $\frac{1}{1-\gamma}$ commonly referred to as the *effective horizon*. Throughout this paper, we primarily focus on the scenario where S, A, B and $\frac{1}{1-\gamma}$ could all be large. Additionally, for notational simplicity, we shall define the vector $P_{s,a,b} \in \mathbb{R}^{1\times S}$ as $P_{s,a,b} \coloneqq P(\cdot \mid s, a, b)$ for any $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$.

Policy, value function, Q-function, and occupancy distribution. Let $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ and $\nu : \mathcal{S} \to \Delta(\mathcal{B})$ be (possibly random) stationary policies of the max-player and the min-player, respectively. In particular, $\mu(\cdot \mid s) \in \Delta(\mathcal{A})$ (resp. $\nu(\cdot \mid s) \in \Delta(\mathcal{B})$) specifies the action selection probability of the max-player (resp. min-player) in state s. The value function $V^{\mu,\nu} : \mathcal{S} \to \mathbb{R}$ for a given product policy $\mu \times \nu$ is defined as

$$V^{\mu,\nu}(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \mid s_0 = s; \mu, \nu\right], \qquad \forall s \in \mathcal{S},$$

where the expectation is taken with respect to the randomness of the trajectory $\{(s_t, a_t, b_t)\}_{t\geq 0}$ induced by the product policy $\mu \times \nu$ (i.e., for any $t \geq 0$, the players take $a_t \sim \mu(\cdot \mid s_t)$ and $b_t \sim \nu(\cdot \mid s_t)$ *independently* conditional on the past) and the probability transition kernel P (i.e., $s_{t+1} \sim P(\cdot \mid s_t, a_t, b_t)$ for $t \geq 0$). Similarly, we can define the Q-function $Q^{\mu,\nu} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ for a given product policy $\mu \times \nu$ as follows

$$Q^{\mu,\nu}(s, a, b) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \mid s_0 = s, a_0 = a, b_0 = b; \mu, \nu\right], \qquad \forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B},$$

where the actions are drawn from $\mu \times \nu$ except for the initial time step (namely, for any $t \geq 1$, we execute $a_t \sim \mu(\cdot \mid s_t)$ and $b_t \sim \nu(\cdot \mid s_t)$ independently conditional on the past). Additionally, for any state distribution $\rho \in \Delta(\mathcal{S})$, we introduce the following notation tailored to the weighted value function of policy pair (μ, ν):

$$V^{\mu,\nu}(\rho) \coloneqq \mathbb{E}_{s\sim\rho}\left[V^{\mu,\nu}(s)\right].$$

Moreover, we define the discounted occupancy measures associated with an initial state distribution $\rho \in \Delta(\mathcal{S})$ and the product policy $\mu \times \nu$ as follows:

$$d^{\mu,\nu}(s; \rho) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid s_0 \sim \rho; \mu, \nu), \qquad \forall s \in \mathcal{S}, \tag{2.1}$$

$$d^{\mu,\nu}(s, a, b; \rho) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a, b_t = b \mid s_0 \sim \rho; \mu, \nu), \qquad \forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}, \tag{2.2}$$

where the sample trajectory $\{(s_t, a_t, b_t)\}_{t\geq 0}$ is initialized with $s_0 \sim \rho$ and then induced by the product policy $\mu \times \nu$ and the transition kernel P as before. It is clearly seen from the above definition that

$$d^{\mu,\nu}(s, a, b; \rho) = d^{\mu,\nu}(s; \rho)\,\mu(a \mid s)\nu(b \mid s). \tag{2.3}$$

Nash equilibrium. In general, the two players have conflicting goals, with the max-player aimed at maximizing the value function and the min-player minimizing the value function. As a result, a standard compromise in Markov games becomes finding a Nash equilibrium (NE). To be precise, a policy pair $(\mu^\star, \nu^\star)$ is said to be a Nash equilibrium if no player can benefit from unilaterally changing her own policy given the opponent's policy (Nash, 1951); that is, for any policies $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ and $\nu : \mathcal{S} \to \Delta(\mathcal{B})$, one has

$$V^{\mu,\nu^\star} \leq V^{\mu^\star,\nu^\star} \leq V^{\mu^\star,\nu}.$$

As is well known (Shapley, 1953), there exists at least one Nash equilibrium $(\mu^\star, \nu^\star)$ in the discounted two-player zero-sum Markov game, and every NE results in the same value function

$$V^\star(s) \coloneqq V^{\mu^\star,\nu^\star}(s) = \max_\mu \min_\nu V^{\mu,\nu}(s) = \min_\nu \max_\mu V^{\mu,\nu}(s).$$

In addition, when the max-player's policy μ is fixed, it is clearly seen that the MG reduces to a single-agent Markov Decision Process (MDP). In light of this, we define, for each $s \in \mathcal{S}$,

$$V^{\mu,\star}(s) \coloneqq \min_\nu V^{\mu,\nu}(s) \qquad \text{and} \qquad V^{\star,\nu} \coloneqq \max_\mu V^{\mu,\nu}(s),$$

each of which corresponds to the optimal value function of one player with the opponent's policy frozen. Moreover, for any policy pair (μ, ν), the following weak duality property always holds:

$$V^{\mu,\star} \leq V^{\mu^\star,\nu^\star} = V^\star \leq V^{\star,\nu}.$$

In this paper, our goal can be posed as calculating a policy pair $(\widehat{\mu}, \widehat{\nu})$ such that

$$V^{\widehat{\mu},\star}(\rho) - \varepsilon \leq V^\star(\rho) \leq V^{\star,\widehat{\nu}}(\rho) + \varepsilon,$$

where $\rho \in \Delta(\mathcal{S})$ is some prescribed initial state distribution, and $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$ denotes the target accuracy level. The gap $V^{\star,\widehat{\nu}}(\rho) - V^{\widehat{\mu},\star}(\rho)$ shall often be referred to as the duality gap of $(\widehat{\mu}, \widehat{\nu})$ in the rest of the present paper.

2.2 Offline dataset (batch dataset)

Suppose that we have access to a historical dataset containing a batch of N sample transitions $\mathcal{D} = \{(s_i, a_i, b_i, s_i')\}_{1 \leq i \leq N}$, which are generated *independently* from a distribution $d_{\mathsf{b}} \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{B})$ and the true probability transition kernel P, namely,

$$(s_i, a_i, b_i) \overset{\text{i.i.d.}}{\sim} d_{\mathsf{b}} \qquad \text{and} \qquad s_i' \overset{\text{ind.}}{\sim} P(\cdot \,|\, s_i, a_i, b_i). \tag{2.4}$$

The goal is to learn an approximate Nash equilibrium on the basis of this historical dataset.

In general, the data distribution d_{b} might deviate from the one generated by a Nash equilibrium $(\mu^\star, \nu^\star)$. As a result, whether reliable learning is feasible depends heavily upon the quality of the historical data. To quantify the quality of the data distribution, Cui and Du (2022b) introduced the following unilateral concentrability condition.

Assumption 1 (Unilateral concentrability)**.** *Suppose that the following quantity*

$$C^\star \coloneqq \max\left\{ \sup_{\mu,s,a,b} \frac{d^{\mu,\nu^\star}(s,a,b;\rho)}{d_{\mathsf{b}}(s,a,b)}, \sup_{\nu,s,a,b} \frac{d^{\mu^\star,\nu}(s,a,b;\rho)}{d_{\mathsf{b}}(s,a,b)} \right\} \tag{2.5}$$

to be finite, where we define $0/0 = 0$ by convention. This quantity $C^\star$ is termed the "unilateral concentrability coefficient."

In words, this quantity $C^\star$ employs certain density ratios to measure the distribution mismatch between the target distribution and the data distribution in hand. On the one hand, Assumption 1 is substantially weaker than the type of uniform coverage requirement (which imposes a uniform bound on the density ratio $\frac{d^{\mu,\nu}(s,a,b;\rho)}{d_{\mathsf{b}}(s,a,b)}$ over all (μ, ν) simultaneously), as (2.5) freezes the policy of one side while exhausting over all policies of the other side. On the other hand, Assumption 1 remains more stringent than a single-policy coverage requirement (which only requires the dataset to cover the part of the state-action space reachable by a given policy pair $(\mu^\star, \nu^\star)$), as (2.5) requires the data to cover those state-action pairs reachable by *any unilateral deviation* from the target policy pair $(\mu^\star, \nu^\star)$. As posited by Cui and Du (2022b), unilateral coverage (i.e., a finite $C^\star < \infty$) is necessary for learning Nash Equilibria in Markov games, which stands in sharp contrast to the single-agent case where single-policy concentrability suffices for finding the optimal policy (Li et al., 2022c; Rashidinejad et al., 2021; Xie et al., 2021).

In this paper, we introduce a modified assumption that might give rise to slightly improved sample complexity bounds.

Assumption 2 (Clipped unilateral concentrability). *Suppose that the following quantity*

$$C^\star_{\mathsf{clipped}} := \max\left\{ \sup_{\mu,s,a,b} \frac{\min\left\{d^{\mu,\nu^\star}(s,a,b;\rho), \frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}(s,a,b)}, \sup_{\nu,s,a,b} \frac{\min\left\{d^{\mu^\star,\nu}(s,a,b;\rho), \frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}(s,a,b)} \right\} \tag{2.6}$$

is finite, where we define $0/0 = 0$ by convention. This quantity $C^\star_{\mathsf{clipped}}$ is termed the "clipped unilateral concentrability coefficient."

In a nutshell, when $d^{\mu,\nu^\star}(s,a,b;\rho)$ or $d^{\mu^\star,\nu}(s,a,b;\rho)$ is reasonably large (i.e., larger than $\frac{1}{S(A+B)}$), Assumption 2 no longer requires the data distribution d_{b} to scale proportionally with $d^{\mu,\nu^\star}$ or $d^{\mu^\star,\nu}$, thus resulting in (slight) relaxation of Assumption 1. Comparing (2.6) with (2.5) immediately reveals that

$$C^\star \geq C^\star_{\mathsf{clipped}}$$

holds all the time. Further, it is straightforward to verify that $C^\star \geq \max\{A, B\}$; in comparison, $C^\star_{\mathsf{clipped}}$ can be as small as $\frac{2AB}{S(A+B)}$, as shown in our lower bound construction in Appendix D.

3 Algorithm and main theory

In this section, we propose a pessimistic model-based offline algorithm — called VI-LCB-Game — to solve the two-player zero-sum Markov games. The proposed algorithm is then shown to achieve minimax-optimal sample complexity in finding an approximate Nash equilibrium of the Markov game given offline data.

3.1 Algorithm design

The empirical Markov game. With the offline dataset $\{(s_i, a_i, b_i, s'_i)\}_{1 \leq i \leq N}$ in hand, we can readily construct an empirical Markov game. To do so, we first compute the sample size

$$N(s,a,b) = \sum_{i=1}^{N} \mathbb{1}\left\{(s_i, a_i, b_i) = (s,a,b)\right\}$$

for each $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. The empirical transition kernel $\widehat{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ is then constructed as follows:

$$\widehat{P}(s' \mid s,a,b) = \begin{cases} \frac{1}{N(s,a,b)} \sum_{i=1}^{N} \mathbb{1}\left\{(s_i, a_i, b_i, s'_i) = (s,a,b,s')\right\}, & \text{if } N(s,a,b) > 0 \\ \frac{1}{S}, & \text{if } N(s,a,b) = 0 \end{cases} \tag{3.1}$$

for any $s' \in \mathcal{S}$ and any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. Throughout this paper, we shall often let $\widehat{P}_{s,a,b} \in \mathbb{R}^{1 \times S}$ abbreviate $\widehat{P}(\cdot \mid s,a,b)$. In addition, the empirical reward function $\widehat{r} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is taken to be

$$\widehat{r}(s,a,b) = \begin{cases} r(s,a,b), & \text{if } N(s,a,b) > 0 \\ 0, & \text{if } N(s,a,b) = 0 \end{cases} \tag{3.2}$$

for any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. Armed with these components, we arrive at an empirical zero-sum Markov game, denoted by $\widehat{\mathcal{MG}} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, \widehat{P}, \widehat{r}, \gamma)$.

Pessimistic Bellman operators. Recall that the classical Bellman operator $\mathcal{T}: \mathbb{R}^{SAB} \to \mathbb{R}^{SAB}$ is defined such that (Lagoudakis and Parr, 2002; Shapley, 1953): for any $Q: \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$,

$$\mathcal{T}(Q)(s,a,b) = r(s,a,b) + \gamma P_{s,a,b} V,$$

where $V: \mathcal{S} \to \mathbb{R}$ is the value function associated with the input Q, i.e.,

$$V(s) \coloneqq \max_{\mu_s \in \Delta(\mathcal{A})} \min_{\nu_s \in \Delta(\mathcal{B})} \mathop{\mathbb{E}}_{a \sim \mu_s, b \sim \nu_s} \big[Q(s,a,b)\big], \qquad \forall s \in \mathcal{S}. \tag{3.3}$$

Note, however, that we are in need of modified versions of the Bellman operator in order to accommodate the offline setting. In this paper, we introduce the pessimistic Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ (resp. $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}$) for the max-player (resp. min-player) as follows: for every $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$,

$$\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q)(s,a,b) \coloneqq \max\left\{\widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V - \beta(s,a,b;V), 0\right\}, \tag{3.4a}$$

$$\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q)(s,a,b) \coloneqq \min\left\{\widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V + \beta(s,a,b;V), \frac{1}{1-\gamma}\right\}, \tag{3.4b}$$

where V is again defined in (3.3). The additional term $\beta(s,a,b;V)$ is incorporated into the operators in order to implement pessimism; informally, we anticipate this penalty term to help $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ (resp. $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}$) produce a conservative estimate of the Q-function from the max-player's (resp. min-player's) viewpoint. Here and throughout, we choose this term based on *Bernstein-style* concentration bounds; specifically, we take

$$\beta(s,a,b;V) = \min\left\{\max\left\{\sqrt{\frac{C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a,b)} \mathsf{Var}_{\widehat{P}_{s,a,b}}(V)}, \frac{2C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)}\right\}, \frac{1}{1-\gamma}\right\} + \frac{4}{N} \tag{3.5}$$

for some sufficiently large constant $C_{\mathsf{b}} > 0$, where $1-\delta$ denotes the target success probability, and the empirical variance term is defined as

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}(V) \coloneqq \widehat{P}_{s,a,b} V^2 - (\widehat{P}_{s,a,b} V)^2. \tag{3.6}$$

It is well known that the classical Bellman operator $\mathcal{T}$ satisfies the γ-contraction property, which guarantees fast global convergence of classical value iteration. As it turns out, the pessimistic Bellman operators introduced above also enjoy the γ-contraction property in the sense that

$$\left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_2)\right\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty \quad \text{and} \quad \left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2)\right\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty; \tag{3.7}$$

see Lemma 1 for precise statements.

Pessimistic value iteration with Bernstein-style penalty. With the pessimistic Bellman operators in place, we are positioned to present the proposed paradigm. Our algorithm maintains the Q-function iterates $\{Q_{\mathsf{pe},t}^{-}\}$, the policy iterates $\{\mu_t^{-}\}$ and $\{\nu_t^{-}\}$, and the value function iterates $\{V_{\mathsf{pe},t}^{-}\}$ from the max-player's perspective; at the same time, it also maintains an analogous group of iterates $\{Q_{\mathsf{pe},t}^{+}\}$, $\{\mu_t^{+}\}$ and $\{\nu_t^{+}\}$ and $\{V_{\mathsf{pe},t}^{+}\}$ from the min-player's perspective. The updates of the two groups of iterates are carried out in a *completely decoupled* manner, except when determining the final output.

In what follows, let us describe the update rules from the max-player's perspective. For notational simplicity, we shall write $\mu(s) \coloneqq \mu(\cdot \mid s) \in \Delta(\mathcal{A})$ and $\nu(s) \coloneqq \nu(\cdot \mid s) \in \Delta(\mathcal{B})$ whenever it is clear from the context. In each round $t = 1, 2, \cdots$, we carry out the following update rules:

1. *Updating Q-function estimates.* Run a pessimistic variant of value iteration to yield

$$Q_{\mathsf{pe},t}^{-} = \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}\left(Q_{\mathsf{pe},t-1}^{-}\right). \tag{3.8}$$

The γ-contraction property (3.7) helps ensure sufficient progress made in each iteration of this update rule.

2. *Updating policy estimates.* We then adjust the policies based on the updated Q-function estimates (3.8). Specifically, for each $s \in \mathcal{S}$, we compute the Nash equilibrium $\big(\mu_t^-(s), \nu_t^-(s)\big) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$ of the zero-sum matrix game with payoff matrix $Q_{\mathsf{pe},t}^-(s,\cdot,\cdot)$. It is worth noting that there is a host of methods for efficiently calculating the NE of a zero-sum matrix game, prominent examples including linear programming and no-regret learning (Freund and Schapire, 1999; Raghavan, 1994; Rakhlin and Sridharan, 2013; Roughgarden, 2016).

3. *Policy evaluation*: for each $s \in \mathcal{S}$, update the value function estimates based on the updated policies $\big(\mu_t^-(s), \nu_t^-(s)\big)$ as follows

$$V_{\mathsf{pe},t}^-(s) = \mathop{\mathbb{E}}_{a\sim\mu_t^-(s), b\sim\nu_t^-(s)}\big[Q_{\mathsf{pe},t}^-(s,a,b)\big].$$

The updates for $\big\{Q_{\mathsf{pe},t}^+\big\}$, $\big\{\mu_t^+\big\}$ and $\big\{\nu_t^+\big\}$ from the min-player's perspective are carried out in an analogous and completely independent manner; see Algorithm 1 for details.

Final output. By running the above update rules for $T = \lceil \frac{\log(N/(1-\gamma))}{\log(1/\gamma)} \rceil$ iterations, we arrive at the Q-function estimates

$$Q_{\mathsf{pe}}^- \coloneqq Q_{\mathsf{pe},T}^- \qquad \text{and} \qquad Q_{\mathsf{pe}}^+ \coloneqq Q_{\mathsf{pe},T}^+, \tag{3.9}$$

in addition to two sets of policy estimates

$$\big(\mu^-, \nu^-\big) \coloneqq \big(\mu_T^-, \nu_T^-\big) \qquad \text{and} \qquad \big(\mu^+, \nu^+\big) \coloneqq \big(\mu_T^+, \nu_T^+\big). \tag{3.10}$$

The final policy estimate of the algorithm is then chosen to be

$$\big(\widehat{\mu}, \widehat{\nu}\big) = \big(\mu^-, \nu^+\big).$$

The full algorithm is summarized in Algorithm 1.

3.2 Theoretical guarantees

Our main result is to uncover the intriguing sample efficiency of the proposed model-based algorithm, as formally stated below.

Theorem 1. *Consider any initial state distribution $\rho \in \Delta(\mathcal{S})$, and suppose that Assumption 2 holds. Assume that $1/2 \leq \gamma < 1$, and consider any $\delta \in (0,1)$ and $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$. Then with probability exceeding $1-\delta$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by Algorithm 1 satisfies*

$$V^{\widehat{\mu},\star}(\rho) - \varepsilon \leq V^\star(\rho) \leq V^{\star,\widehat{\nu}}(\rho) + \varepsilon,$$

as long as the sample size exceeds

$$N \geq c_1 \frac{C_{\mathsf{clipped}}^\star S(A+B)}{(1-\gamma)^3 \varepsilon^2} \log \frac{N}{(1-\gamma)\delta}$$

for some sufficiently large constant $c_1 > 0$.

The sample complexity needed for Algorithm 1 to compute a policy pair with ε-duality gap is at most

$$\widetilde{O}\left(\frac{C_{\mathsf{clipped}}^\star S(A+B)}{(1-\gamma)^3 \varepsilon^2}\right), \tag{3.11}$$

which accommodates any target accuracy within the range $\big(0, \frac{1}{1-\gamma}\big]$. In addition to linear dependency on $C_{\mathsf{clipped}}^\star$, the sample complexity bound (3.11) scales linearly (as opposed to quadratically) with the aggregate size $A + B$ of the individual action spaces. It is noteworthy that our algorithm is a fairly straightforward implementation of the model-based approach (except that the pessimism principle is incorporated), and does not require either sample splitting or sophisticated schemes like variance reduction (Li et al., 2021; Xie et al., 2021; Yan et al., 2022; Zhang et al., 2020a,b).

As it turns out, the above sample complexity theory for Algorithm 1 matches the minimax lower limit modulo some logarithmic term, as asserted by the following theorem. The proof is postponed to Appendix D.

Algorithm 1 Value iteration with lower confidence bounds for zero-sum Markov games (VI-LCB-Game).

Initialization: set $Q^{-}_{\mathsf{pe},0}(s,a,b) = 0$ and $Q^{+}_{\mathsf{pe},0}(s,a,b) = \frac{1}{1-\gamma}$ for all $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$; set $T = \lceil \frac{\log(N/(1-\gamma))}{\log(1/\gamma)} \rceil$.

Compute the empirical transition kernel $\widehat{P}$ as (3.1) and the empirical reward function $\widehat{r}$ as (3.2).

For: $t = 1,\dots,T$ **do**

- Update

$$Q^{-}_{\mathsf{pe},t}(s,a,b) = \widehat{\mathcal{T}}^{-}_{\mathsf{pe}}\left(Q^{-}_{\mathsf{pe},t-1}\right) = \max\left\{\widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b}V^{-}_{\mathsf{pe},t-1} - \beta\left(s,a,b;V^{-}_{\mathsf{pe},t-1}\right), 0\right\},$$
$$Q^{+}_{\mathsf{pe},t}(s,a,b) = \widehat{\mathcal{T}}^{+}_{\mathsf{pe}}\left(Q^{+}_{\mathsf{pe},t-1}\right) = \min\left\{\widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b}V^{+}_{\mathsf{pe},t-1} + \beta\left(s,a,b;V^{+}_{\mathsf{pe},t-1}\right), \frac{1}{1-\gamma}\right\},$$

  where

$$\beta(s,a,b;V) = \min\left\{\max\left\{\sqrt{\frac{C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a,b)}\mathsf{Var}_{\widehat{P}_{s,a,b}}(V)}, \frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)}\right\}, \frac{1}{1-\gamma}\right\} + \frac{4}{N}$$

  for some sufficiently large constant $C_{\mathsf{b}} > 0$, with $\mathsf{Var}_{\widehat{P}_{s,a,b}}(V)$ defined in (3.6).

- For each $s \in \mathcal{S}$, compute

$$\left(\mu^{-}_{t}(s), \nu^{-}_{t}(s)\right) = \mathsf{MatrixNash}\left(Q^{-}_{\mathsf{pe},t}(s,\cdot,\cdot)\right),$$
$$\left(\mu^{+}_{t}(s), \nu^{+}_{t}(s)\right) = \mathsf{MatrixNash}\left(Q^{+}_{\mathsf{pe},t}(s,\cdot,\cdot)\right),$$

  where for any matrix $M \in \mathbb{R}^{A\times B}$, the function $\mathsf{MatrixNash}(M)$ returns a solution $(\widehat{w},\widehat{z})$ to the minimax program $\max_{w\in\Delta(\mathcal{A})}\min_{z\in\Delta(\mathcal{B})} w^{\top}Mz$.

- For each $s \in \mathcal{S}$, update

$$V^{-}_{\mathsf{pe},t}(s) = \mathbb{E}_{a\sim\mu^{-}_{t}(s), b\sim\nu^{-}_{t}(s)}\left[Q^{-}_{\mathsf{pe},t}(s,a,b)\right],$$
$$V^{+}_{\mathsf{pe},t}(s) = \mathbb{E}_{a\sim\mu^{+}_{t}(s), b\sim\nu^{+}_{t}(s)}\left[Q^{+}_{\mathsf{pe},t}(s,a,b)\right].$$

Output: the policy pair $(\widehat{\mu},\widehat{\nu})$, where $\widehat{\mu} = \{\mu^{-}_{T}(s)\}_{s\in\mathcal{S}}$ and $\widehat{\nu} = \{\nu^{+}_{T}(s)\}_{s\in\mathcal{S}}$.

Theorem 2. *Consider any $S \geq 2$, $A \geq 2$, $B \geq 2$, $\gamma \in [\frac{2}{3}, 1)$ and $C^{\star}_{\mathsf{clipped}} \geq \frac{2AB}{S(A+B)}$, and define the set*

$$\mathsf{MG}\left(C^{\star}_{\mathsf{clipped}}\right) := \left\{\{\mathcal{MG},\rho,d_{\mathsf{b}}\}\ \middle|\ |\mathcal{S}| = S,\ |\mathcal{A}| = A,\ |\mathcal{B}| = B,\ \rho\in\Delta(\mathcal{S}),\ d_{\mathsf{b}}\in\Delta(\mathcal{S}\times\mathcal{A}\times\mathcal{B}),\ \exists \text{ an NE } (\mu^{\star},\nu^{\star}) \text{ of } \mathcal{MG} \text{ such that } \max\left\{\sup_{\mu,s,a,b}\frac{\min\left\{d^{\mu,\nu^{\star}}(s,a,b;\rho), \frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}(s,a,b)}, \sup_{\nu,s,a,b}\frac{\min\left\{d^{\mu^{\star},\nu}(s,a,b;\rho), \frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}(s,a,b)}\right\} = C^{\star}_{\mathsf{clipped}}\right\}.$$

Then there exist some universal constants $c_2, c_{\varepsilon} > 0$ such that: for any $\varepsilon \in (0, \frac{1}{c_{\varepsilon}(1-\gamma)\log(A+B)}]$, if the sample size obeys

$$N < \frac{c_2 S(A+B)C^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2\log(A+B)},$$

then one necessarily has

$$\inf_{(\widehat{\mu},\widehat{\nu})}\sup_{\{\mathcal{MG},\rho,d_{\mathsf{b}}\}\in\mathsf{MG}(C^{\star}_{\mathsf{clipped}})}\mathbb{E}\left[V^{\star,\widehat{\nu}}(\rho) - V^{\widehat{\mu},\star}(\rho)\right] \geq \varepsilon.$$

Here, the infimum is taken over all estimators $(\widehat{\mu}, \widehat{\nu})$ for the Nash equilibrium based on the batch dataset $\mathcal{D} = \{(s_i, a_i, b_i, s_i')\}_{i=1}^n$ generated according to (2.4).

As a direct implication of Theorem 2, if the total number of samples in the offline dataset obeys

$$N < \frac{c_2 S\left(A+B\right) C^{\star}_{\mathsf{clipped}}}{\left(1-\gamma\right)^3 \varepsilon^2 \log\left(A+B\right)},$$

then one can construct a hard Markov game instance such that no algorithm whatsoever can reach a duality gap below ε. This taken collectively (3.11) unveils, up to some logarithmic factor, the minimax statistical limit for finding NEs based on offline data.

Our theory makes remarkable improvement upon prior art, which can be seen through comparisons with the most relevant prior work Cui and Du (2022b) (even though the focus therein is finite-horizon zero-sum MGs). On a high level, Cui and Du (2022b) proposed an algorithm that combines pessimistic value iteration with variance reduction (also called reference-advantage decomposition (Zhang et al., 2020b)), which provably finds an ε-Nash policy pair using

$$\widetilde{O}\left(\frac{C^{\star} SABH^3}{\varepsilon^2}\right) \tag{3.12}$$

sample trajectories, provided that $\varepsilon \leq 1/H$. Here, H stands for the horizon length of the finite-horizon Markov game, and $C^{\star}$ is the unilateral concentrability coefficient tailored to the finite-horizon setting. Despite the difference between discounted infinite-horizon and finite-horizon settings, our algorithm design and theory achieve several improvements upon Cui and Du (2022b).

- Perhaps most importantly, our result scales linearly in the total number of individual actions $A + B$ (as opposed to the number of joint actions AB as in Cui and Du (2022b)), which manages to alleviate the curse of multiple agents.
- Our theory accommodates the full ε-range $\left(0, \frac{1}{1-\gamma}\right]$, which is much wider than the range $(0, 1/H]$ covered by Cui and Du (2022b) (if we view the effective horizon $\frac{1}{1-\gamma}$ in the infinite-horizon case and the horizon length H in the finite-horizon counterpart as equivalent).
- The algorithm design herein is substantially simpler than Cui and Du (2022b): neither does it require sample splitting to decouple statistical dependency, nor does it rely on reference-advantage decomposition techniques to sharpen the horizon dependency.

While we were finalizing the present manuscript, we became aware of the independent work Cui and Du (2022a) proposing a different offline algorithm — based on incorporation of strategy-wise lower confidence bounds — that improved the prior art as well. When it comes to two-player zero-sum Markov games with finite horizon and non-stationary transition kernels, Cui and Du (2022a, Algorithm 1) provably yields an ε-Nash policy pair using

$$\widetilde{O}\left(\frac{C^{\star} S\left(A+B\right) H^4}{\varepsilon^2}\right) \tag{3.13}$$

sample trajectories each containing H samples. This bound (3.13) is *at least* a factor of H above the minimax limit. It is worth noting that Cui and Du (2022a) is able to accommodate offline multi-agent general-sum MGs, although the algorithm proposed therein becomes computationally intractable when going beyond two-player zero-sum MGs.

4 Related works

Offline RL and pessimism principle. The principle of pessimism in the face of uncertainty, namely, being conservative in value estimation of those state-action pairs that have been under-covered, has been adopted extensively in recent development of offline RL. A highly incomplete list includes Chang et al. (2021); Cui and Du (2022b); Jin et al. (2021b); Kidambi et al. (2020); Kumar et al. (2020); Li et al. (2022b,c); Liu et al. (2020); Lu et al. (2022); Munos (2003, 2007); Rashidinejad et al. (2021); Shi et al. (2022); Uehara and

Sun (2021); Wang et al. (2022); Xie et al. (2021); Yan et al. (2022); Yin et al. (2021a,c); Yin and Wang (2021); Yu et al. (2021a,b, 2020); Zanette et al. (2021); Zhang et al. (2021c); Zhong et al. (2022), which unveiled the efficacy of the pessimism principle in both model-based and model-free approaches. Among this body of prior works, the ones that are most related to the current paper are Cui and Du (2022a,b); Zhong et al. (2022), both of which focused on episodic finite-horizon zero-sum Markov games with two players. More specifically, Cui and Du (2022b) demonstrated that a unilateral concentrability condition is necessary for learning NEs in offline settings, and proposed a pessimistic value iteration with reference-advantage decomposition to enable sample efficiency; Zhong et al. (2022) proposed a *Pessimistic Minimax Value Iteration* algorithm which achieves appealing sample complexity in the presence of linear function representation, which was recently improved by Xiong et al. (2022). The concurrent work Cui and Du (2022a) proposed a different pessimistic algorithm that designed LCBs for policy pairs instead of state-action pairs; for two-player zero-sum MGs, their algorithm is capable of achieving a sample complexity proportional to $A+B$. In the single-agent offline RL setting, Li et al. (2022c); Rashidinejad et al. (2021); Yan et al. (2022) studied offline RL for infinite-horizon MDPs, and Jin et al. (2021b); Li et al. (2022c); Shi et al. (2022); Xie et al. (2021) looked at the finite-horizon episodic counterpart, all of which operate upon some single-policy concentrability assumptions. Among these works, Li et al. (2022c) and Yan et al. (2022) achieved minimax-optimal sample complexity $\widetilde{O}(\frac{SC^{\star}}{(1-\gamma)^3\varepsilon^2})$ for discounted infinite-horizon MDPs by means of model-based and model-free algorithms, respectively; similar results have been established for finite-horizon MDPs as well (Li et al., 2022c; Shi et al., 2022; Xie et al., 2021; Yin et al., 2021b,c).

Multi-agent RL and Markov games. The concept of Markov games — also under the name of stochastic games — dated back to Shapley (1953), which has become a central framework to model competitive multi-agent decision making. A large strand of prior works studied how to efficiently solve Markov games when perfect model description is available (Cen et al., 2021; Chen et al., 2021b; Daskalakis et al., 2020, 2022; Hansen et al., 2013; Hu and Wellman, 2003; Littman, 1994; Littman et al., 2001; Mao and Başar, 2022; Perolat et al., 2015; Wei et al., 2021; Zhao et al., 2021). Recent years have witnessed much activity in studying the sample efficiency of learning Nash equilibria in zero-sum Markov games, covering multiple different types of sampling schemes; for instance, Bai and Jin (2020); Bai et al. (2020); Chen et al. (2021b); Daskalakis et al. (2022); Jin et al. (2021a); Liu et al. (2021); Mao and Başar (2022); Song et al. (2021); Tian et al. (2021); Wei et al. (2017); Xie et al. (2020) focused on the online explorative environments, whereas Zhang et al. (2020a) paid attention to the scenario that assumes sampling access to a generative model. While the majority of these works exhibited a sample complexity that scales at least as $\widetilde{O}(SAB)$ in order to learn an approximate NE, the recent work Bai et al. (2020) proposed a V-learning algorithm attaining a sample complexity that scales linearly with $S(A+B)$, thus matching the minimax-optimal lower bound up to a factor of H^2. When a generative model is available, Li et al. (2022a) further developed an algorithm that learns ε-Nash using $\widetilde{O}\big(\frac{H^4S(A+B)}{\varepsilon^2}\big)$ samples, which attains the minimax lower bound for non-stationary finite-horizon MGs. The setting of general-sum multi-player Markov games is much more challenging, given that learning Nash equilibria is known to be PPAD-complete (Daskalakis, 2013; Daskalakis et al., 2009). Shifting attention to more tractable solution concepts, Daskalakis et al. (2022); Jin et al. (2021a); Mao and Başar (2022); Song et al. (2021) proposed algorithms that provably learn (coarse) correlated equilibria with sample complexities that scale linearly with $\max_i A_i$ (where A_i is the number of actions of the i-th player), thereby breaking the curse of multi-agents. Additionally, there have also been several works investigating the turn-based setting where the two players take actions in turn; see Cui and Yang (2021); Jia et al. (2019); Jin et al. (2022); Sidford et al. (2020). Moreover, another two works Abe and Kaneko (2020); Zhang et al. (2021b) studied offline sampling oracles under uniform coverage requirements (which are clearly more stringent than the unilateral concentrability assumption). The interested readers are also referred to Yang and Wang (2020); Zhang et al. (2021a) for an overview of recent development.

Model-based RL. The method proposed in the current paper falls under the category of model-based algorithms, which decouple model estimation and policy learning (planning). The model-based approach has been extensively studied in the single-agent setting including the online exploration setting (Azar et al., 2017), the case with a generative model (Agarwal et al., 2020; Azar et al., 2013; Jin and Sidford, 2021; Li et al., 2020; Wang et al., 2021), the offline RL setting (Li et al., 2022c; Xie et al., 2021), and turn-based Markov games (Cui and Yang, 2021). Encouragingly, the model-based approach is capable of attaining minimax-optimal sample complexities in a variety of settings (e.g., Agarwal et al. (2020); Azar et al. (2017); Li et al. (2022c)), sometimes even without incurring any burn-in cost (Cui and Yang, 2021; Li et al., 2022c, 2020). The method proposed in Cui and Du (2022b) also exhibited the flavor of a model-based algorithm, although an additional variance reduction scheme is incorporated in order to optimize the horizon dependency.

5 Discussion

In the present paper, we have proposed a model-based offline algorithm, which leverages the principle of pessimism in solving two-player zero-sum Markov games on the basis of past data. In order to find an ε-approximate Nash equilibrium of the Markov game, our algorithm requires no more than $\widetilde{O}\big(\frac{S(A+B)C^{\star}}{(1-\gamma)^3\varepsilon^2}\big)$ samples, and this sample complexity bound is provably minimax optimal for the entire range of target accuracy level $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$. Our theory has improved upon prior sample complexity bounds in Cui and Yang (2021) in terms of the dependency on the size of the action space. Another appealing feature is the simplicity of our algorithm, which does not require complicated variance reduction schemes and is hence easier to implement and interpret. Moving forward, there are a couple of interesting directions that are worthy of future investigation. For instance, one natural extension is to explore whether the current algorithmic idea and analysis extend to multi-agent general-sum Markov games, with the goal of learning other solutions concepts of equilibria like coarse correlated equilibria (given that finding Nash equilibria in general-sum games is PPAD-complete). Another topic of interest is to design model-free algorithms for offline NE learning in zero-sum or general-sum Markov games. Furthermore, the current paper focuses attention on tabular Markov games, and it would be of great interest to design sample-efficient offline multi-agent algorithms in the presence of function approximation.

Acknowledgements

Y. Yan is supported in part by the Charlotte Elizabeth Procter Honorific Fellowship. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grants CCF-2221009, CCF-1907661, IIS-2218713, DMS-2014279, and IIS-2218773. J. Fan is supported in part by the NSF grants DMS-1712591, DMS-2052926, DMS-2053832, and DMS-2210833. Y. Chen thanks Shicong Cen for helpful discussions about Markov games.

A Additional notation

Let us collect a set of additional notation that will be used in the analysis. First of all, for any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, any vector $V \in \mathbb{R}^S$ and any probability transition kernel $P: \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \Delta(\mathcal{S})$, we define

$$\mathsf{Var}_{P_{s,a,b}}(V) = P_{s,a,b}\big(V \circ V\big) - \big(P_{s,a,b}V\big)^2, \tag{A.1}$$

where $P_{s,a,b}$ abbreviates $P(\cdot \,|\, s,a,b)$ as usual. We use the standard notation $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ to denote $|f(n)| \leq Cg(n)$ for some universal constant $C>0$ when n is sufficiently large; we let $f(n) \gtrsim g(n)$ denote $f(n) \geq C|g(n)|$ for some constant $C>0$ when n is large enough; and we employ $f(n) \asymp g(n)$ to indicate that $f(n) \gtrsim g(n)$ and $f(n) \lesssim g(n)$ hold simultaneously.

When the max-player's policy μ is fixed, the Markov game reduces to a (single-agent) MDP for the min-player. For any MDP, it is known that there exists at least one policy that simultaneously maximizes the value function (resp. Q-function) for all states (resp. state-action pairs) (Bertsekas, 2017). In light of this, when the policy μ of the max-player is frozen, we denote by $\nu_{\mathsf{br}}(\mu)$ the optimal policy of the min-player, which shall often be referred to as the best response of the min-player when the max-player adopts policy μ. Similarly, we can define the best response of the max-player when the min-player adopts policy ν, which we denoted by $\mu_{\mathsf{br}}(\nu)$. These allow one to define

$$V^{\mu,\star}(s) \coloneqq V^{\mu,\nu_{\mathsf{br}}(\mu)}(s) = \min_{\nu} V^{\mu,\nu}(s), \qquad V^{\star,\nu}(s) \coloneqq V^{\mu_{\mathsf{br}}(\nu),\nu}(s) = \max_{\mu} V^{\mu,\nu}(s)$$

for all $s \in \mathcal{S}$, and

$$Q^{\mu,\star}(s,a,b) := Q^{\mu,\nu_{\mathsf{br}}(\mu)}(s,a,b) = \min_{\nu} Q^{\mu,\nu}(s,a,b), \quad Q^{\star,\nu}(s,a,b) := Q^{\mu_{\mathsf{br}}(\nu),\nu}(s,a,b) = \max_{\mu} Q^{\mu,\nu}(s,a,b)$$

for all $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. Note that the definitions of $V^{\mu,\star}$ and $V^{\star,\nu}$ here are consistent with the ones in Section 2.

B Proof of Theorem 1

Towards proving Theorem 1, we first state a slightly stronger result as follows.

Theorem 3. *Consider any initial state distribution $\rho \in \Delta(\mathcal{S})$, and suppose that Assumption 2 holds. Assume that $1/2 \le \gamma < 1$. Then with probability exceeding $1-\delta$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by Algorithm 1 satisfies*

$$V^{\star}(\rho) - V^{\widehat{\mu},\star}(\rho) \lesssim \sqrt{\frac{C^{\star}_{\mathsf{clipped}} S(A+B)}{(1-\gamma)^3 N} \log \frac{N}{(1-\gamma)\delta}} + \frac{C^{\star}_{\mathsf{clipped}} S(A+B)}{(1-\gamma)^2 N} \log \frac{N}{(1-\gamma)\delta}, \tag{B.1a}$$

$$V^{\star,\widehat{\nu}}(\rho) - V^{\star}(\rho) \lesssim \sqrt{\frac{C^{\star}_{\mathsf{clipped}} S(A+B)}{(1-\gamma)^3 N} \log \frac{N}{(1-\gamma)\delta}} + \frac{C^{\star}_{\mathsf{clipped}} S(A+B)}{(1-\gamma)^2 N} \log \frac{N}{(1-\gamma)\delta}. \tag{B.1b}$$

As an immediate consequence, the duality gap of $(\widehat{\mu}, \widehat{\nu})$ obeys, with probability at least $1-\delta$, that

$$V^{\star,\widehat{\nu}}(\rho) - V^{\widehat{\mu},\star}(\rho) \lesssim \sqrt{\frac{C^{\star}_{\mathsf{clipped}} S(A+B)}{(1-\gamma)^3 N} \log \frac{N}{(1-\gamma)\delta}} + \frac{C^{\star}_{\mathsf{clipped}} S(A+B)}{(1-\gamma)^2 N} \log \frac{N}{(1-\gamma)\delta}. \tag{B.2}$$

As can be straightforwardly verified, Theorem 1 is a direct consequence of Theorem 3 (by taking the right-hand side of (B.2) to be no larger than ε). The remainder of this section is thus dedicated to establishing Theorem 3.

B.1 Preliminary facts

Before continuing, we collect several preliminary facts that will be useful throughout.

1. For any $Q_1, Q_2 : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, we have

   $$\|V_1 - V_2\|_\infty \le \|Q_1 - Q_2\|_\infty, \tag{B.3}$$

   where V_1 (resp. V_2) denotes the value function associated with Q_1 (resp. Q_2); see (3.3) for the precise definition.

2. For any $V_1, V_2 : \mathcal{S} \to \left[0, \frac{1}{1-\gamma}\right]$, any probability transition kernel $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ and any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, we have

   $$\left|\mathsf{Var}_{P_{s,a,b}}(V_1) - \mathsf{Var}_{P_{s,a,b}}(V_2)\right| \le \frac{4}{1-\gamma} \|V_1 - V_2\|_\infty, \tag{B.4}$$

   where $\mathsf{Var}_{P,s,a,b}(V)$ is defined in (A.1).

3. As a consequence, we also know that for any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and any $V_1, V_2 : \mathcal{S} \to \left[0, \frac{1}{1-\gamma}\right]$, the corresponding penalty terms (cf. (3.5)) obey

   $$|\beta(s,a,b;V_1) - \beta(s,a,b;V_2)| \le 2\|V_1 - V_2\|_\infty. \tag{B.5}$$

The proof of the preceding results can be found in Appendix C.1.

B.2 Step 1: key properties of pessimistic Bellman operators

Recall the definition of the pessimistic Bellman operators $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ and $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}$ introduced in (3.4). The following lemma gathers a couple of key properties of these two operators.

Lemma 1. *The following properties hold true:*

- *(Monotonicity) For any $Q_1 \geq Q_2$ we have $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_1) \geq \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_2)$ and $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) \geq \widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2)$;*
- *(Contraction) Both operators are γ-contractive in the ℓ_∞ sense, i.e.,*

$$\big\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_2)\big\|_\infty \leq \gamma\big\|Q_1 - Q_2\big\|_\infty, \qquad \big\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2)\big\|_\infty \leq \gamma\left\|Q_1 - Q_2\right\|_\infty$$

  for any Q_1 and Q_2;
- *(Uniqueness of fixed points) $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ (resp. $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}$) has a unique fixed point $Q_{\mathsf{pe}}^{-\star}$ (resp. $Q_{\mathsf{pe}}^{+\star}$), which also satisfies $0 \leq Q_{\mathsf{pe}}^{\star-}(s,a,b) \leq \frac{1}{1-\gamma}$ (resp. $0 \leq Q_{\mathsf{pe}}^{+\star}(s,a,b) \leq \frac{1}{1-\gamma}$) for any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$.*

Proof. See Appendix C.2. □

Next, we make note of several immediate consequences of Lemma 1. Here and throughout, $V_{\mathsf{pe}}^{-\star}$ and $V_{\mathsf{pe}}^{+\star}$ are defined to be the value functions (see (3.3)) associated with $Q_{\mathsf{pe}}^{-\star}$ and $Q_{\mathsf{pe}}^{+\star}$, respectively.

- First of all, the above lemma implies that

$$Q_{\mathsf{pe},t}^{-} \leq Q_{\mathsf{pe}}^{-\star} \quad (\forall t \geq 0) \qquad \text{and hence} \qquad Q_{\mathsf{pe}}^{-} \leq Q_{\mathsf{pe}}^{-\star}. \tag{B.6}$$

  To see this, we first note that $Q_{\mathsf{pe},0}^{-} = 0 \leq Q_{\mathsf{pe}}^{-\star}$. Next, suppose that $Q_{\mathsf{pe},t}^{-} \leq Q_{\mathsf{pe}}^{-\star}$ for some iteration $t \geq 0$, then the monotonicity of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ (cf. Lemma 1) tells us that

$$Q_{\mathsf{pe},t+1}^{-} = \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}\big(Q_{\mathsf{pe},t}^{-}\big) \leq \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_{\mathsf{pe}}^{-\star}) = Q_{\mathsf{pe}}^{-\star},$$

  from which (B.6) follows.
- In addition, the γ-contraction property in Lemma 1 leads to

$$\big\|V_{\mathsf{pe}}^{-} - V_{\mathsf{pe}}^{-\star}\big\|_\infty \leq \big\|Q_{\mathsf{pe}}^{-} - Q_{\mathsf{pe}}^{-\star}\big\|_\infty \leq \frac{1}{N}, \tag{B.7}$$

  To justify this, observe that

$$\begin{aligned}\big\|Q_{\mathsf{pe},t}^{-} - Q_{\mathsf{pe}}^{-\star}\big\|_\infty &= \Big\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_{\mathsf{pe},t-1}^{-}) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q_{\mathsf{pe}}^{-\star})\Big\|_\infty \leq \gamma\left\|Q_{\mathsf{pe},t-1}^{-} - Q_{\mathsf{pe}}^{-\star}\right\|_\infty \\ &\leq \cdots \leq \gamma^t\left\|Q_{\mathsf{pe},0}^{-} - Q_{\mathsf{pe}}^{-\star}\right\|_\infty \leq \frac{\gamma^t}{1-\gamma},\end{aligned}$$

  which together with $T = \lceil \frac{\log(N/(1-\gamma))}{\log(1/\gamma)} \rceil$ and (B.3) gives

$$\big\|V_{\mathsf{pe}}^{-} - V_{\mathsf{pe}}^{-\star}\big\|_\infty \leq \big\|Q_{\mathsf{pe}}^{-} - Q_{\mathsf{pe}}^{-\star}\big\|_\infty = \big\|Q_{\mathsf{pe},T}^{-} - Q_{\mathsf{pe}}^{-\star}\big\|_\infty \leq \frac{\gamma^T}{1-\gamma} \leq \frac{1}{N}.$$

- A similar argument also yields

$$Q_{\mathsf{pe}}^{+} \geq Q_{\mathsf{pe}}^{+\star}, \qquad \big\|Q_{\mathsf{pe}}^{+} - Q_{\mathsf{pe}}^{+\star}\big\|_\infty \leq 1/N, \qquad \big\|V_{\mathsf{pe}}^{+} - V_{\mathsf{pe}}^{+\star}\big\|_\infty \leq 1/N. \tag{B.8}$$

B.3 Step 2: decoupling statistical dependency and establishing pessimism

To proceed, we rely on the following theorem to quantify the difference between $\widehat{P}$ and P when projected onto a value function direction.

Lemma 2. *For any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ satisfying $N(s,a,b)\geq 1$, with probability exceeding $1-\delta$,*

$$\left|\big(\widehat{P}_{s,a,b}-P_{s,a,b}\big)\widetilde{V}\right| \lesssim \sqrt{\frac{1}{N(s,a,b)}\mathsf{Var}_{\widehat{P}_{s,a,b}}\big(\widetilde{V}\big)\log\frac{N}{(1-\gamma)\,\delta}} + \frac{\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)\,N(s,a,b)} \tag{B.9}$$

and

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}\big(\widetilde{V}\big) \leq 2\mathsf{Var}_{P_{s,a,b}}\big(\widetilde{V}\big) + O\left(\frac{1}{(1-\gamma)^2 N(s,a,b)}\log\frac{N}{(1-\gamma)\,\delta}\right) \tag{B.10}$$

hold simultaneously for all $\widetilde{V}\in\mathbb{R}^S$ satisfying $0\leq\widetilde{V}\leq\frac{1}{1-\gamma}1$ and $\min\left\{\|\widetilde{V}-V_{\mathsf{pe}}^{-\star}\|_\infty, \|\widetilde{V}-V_{\mathsf{pe}}^{+\star}\|_\infty\right\}\leq 1/N$.

In words, the first result (B.9) delivers some Bernstein-type concentration bound, whereas the second result (B.10) guarantees that the empirical variance estimate (i.e., the plug-in estimate) is close to the true variance. It is worth noting that Lemma 2 does *not* require $\widetilde{V}$ to be statistically independent from $\widehat{P}_{s,a,b}$, which is particularly crucial when coping with complicated statistical dependency of our iterative algorithm. The proof of Lemma 2 is established upon a leave-one-out analysis argument (see, e.g., Agarwal et al. (2020); Chen et al. (2021a); Li et al. (2022c, 2020); Ma et al. (2020)) that helps decouple statistical dependency. Armed with Lemma 2, we can readily see that

$$\left|\big(\widehat{P}_{s,a,b}-P_{s,a,b}\big)\widetilde{V}\right| + \frac{4}{N} \leq \beta\big(s,a,b;\widetilde{V}\big) \tag{B.11}$$

holds for any $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ satisfying $N(s,a,b)\geq 1$ and any $\widetilde{V}$ satisfying the conditions in Lemma 2. In turn, this important fact allows one to justify that Q_{pe}^{-} (resp. Q_{pe}^{+}) is indeed an upper (resp. lower) bound on $Q^{\widehat{\mu},\star}$ (resp. $Q^{\star,\widehat{\nu}}$), as formalized below.

Lemma 3. *With probability exceeding $1-\delta$, it holds that*

$$Q_{\mathsf{pe}}^{-}\leq Q^{\widehat{\mu},\star}, \qquad Q_{\mathsf{pe}}^{+}\geq Q^{\star,\widehat{\nu}}, \qquad V_{\mathsf{pe}}^{-}\leq V^{\widehat{\mu},\star} \qquad \textit{and} \qquad V_{\mathsf{pe}}^{+}\geq V^{\star,\widehat{\nu}}.$$

Proof. See Appendix C.4. □

This lemma makes clear a key rationale for the principle of pessimism: we would like the Q-function estimates to be always conservative uniformly over all entries.

B.4 Step 3: bounding $V^\star(\rho)-V^{\widehat{\mu},\star}(\rho)$ and $V^{\star,\widehat{\nu}}(\rho)-V^\star(\rho)$

Before proceeding to bound $V^\star - V^{\widehat{\mu},\star}$, we first develop a lower bound on V_{pe}^{-}, given that $V^{\widehat{\mu},\star}$ is lower bounded by V_{pe}^{-} (according to Lemma 3). Towards this end, we invoke the definition of V_{pe}^{-} to reach

$$V_{\mathsf{pe}}^{-}(s) = \max_{\mu(s)\in\Delta(\mathcal{A})}\min_{\nu(s)\in\Delta(\mathcal{B})}\mathop{\mathbb{E}}_{a\sim\mu(s),b\sim\nu(s)}\left[Q_{\mathsf{pe}}^{-}(s,\cdot,\cdot)\right] \geq \min_{\nu(s)\in\Delta(\mathcal{B})}\mathop{\mathbb{E}}_{a\sim\mu^\star(s),b\sim\nu(s)}\left[Q_{\mathsf{pe}}^{-}(s,a,b)\right], \tag{B.12}$$

where we set the policy of the max-player to be $\mu^\star$ on the right-hand side of the above equation. Clearly, there exists a deterministic policy $\nu_0:\mathcal{S}\to\Delta(\mathcal{B})$ such that

$$\nu_0(s) = \arg\min_{\nu(s)\in\Delta(\mathcal{B})}\mathop{\mathbb{E}}_{a\sim\mu^\star(s),b\sim\nu(s)}\left[Q_{\mathsf{pe}}^{-}(s,a,b)\right] \tag{B.13}$$

for any $s\in\mathcal{S}$; for instance, one can simply set, for any $s\in\mathcal{S}$,

$$\nu_0(s)=\mathbb{1}_{b_s} \qquad \text{with } b_s \coloneqq \arg\max_{b\in\mathcal{B}}\ \big\langle \mu^\star(s), Q_{\mathsf{pe}}^{-}(s,\cdot,b)\big\rangle, \tag{B.14}$$

with $\mathbb{1}_{b_s}$ denoting a probability vector that is nonzero only in b_s. This deterministic policy ν_0 helps us lower bound V_{pe}^-, as accomplished in the following lemma. Here and below, we define two vectors $r^{\mu^\star,\nu_0}, \beta^{\mu^\star,\nu_0} \in \mathbb{R}^S$ and a probability transition kernel $P^{\mu^\star,\nu_0}: \mathcal{S} \to \Delta(\mathcal{S})$ restricted to $\mu^\star$ and ν_0 such that: for any $s, s' \in \mathcal{S}$,

$$r^{\mu^\star,\nu_0}(s) := \mathbb{E}_{a\sim\mu^\star(s), b\sim\nu_0(s)}\left[r(s,a,b)\right], \tag{B.15a}$$

$$\beta^{\mu^\star,\nu_0}(s) := \mathbb{E}_{a\sim\mu^\star(s), b\sim\nu_0(s)}\left[\beta\left(s,a,b;V_{\mathsf{pe}}^-\right)\right], \tag{B.15b}$$

$$P^{\mu^\star,\nu_0}(s' \mid s) := \mathbb{E}_{a\sim\mu^\star(s), b\sim\nu_0(s)}\left[P(s' \mid s,a,b)\right]. \tag{B.15c}$$

Lemma 4. *With probability exceeding $1-\delta$, we have*

$$V_{\mathsf{pe}}^- \geq r^{\mu^\star,\nu_0} + \gamma P^{\mu^\star,\nu_0} V_{\mathsf{pe}}^- - 2\beta^{\mu^\star,\nu_0}. \tag{B.16}$$

Proof. See Appendix C.5. □

In addition, we can invoke Lemma 3 and the fact that $V^\star = V^{\mu^\star,\nu^\star} = V^{\mu^\star,\star}$ to reach

$$V^\star - V^{\widehat{\mu},\star} = V^{\mu^\star,\star} - V^{\widehat{\mu},\star} \leq V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-, \tag{B.17}$$

which motivates us to look at $V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-$. Towards this, we note that the Bellman equation tells us that

$$V^{\mu^\star,\nu_0} = r^{\mu^\star,\nu_0} + \gamma P^{\mu^\star,\nu_0} V^{\mu^\star,\nu_0}. \tag{B.18}$$

Taking (B.16) and (B.18) collectively yields

$$V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^- \leq \gamma P^{\mu^\star,\nu_0}\left(V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-\right) + 2\beta^{\mu^\star,\nu_0}, \tag{B.19}$$

thus resulting in a "self-bounding" type of relations. Applying (B.19) recursively, we arrive at

$$\begin{aligned}
\rho^\top\left(V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-\right) &\leq \gamma\rho^\top P^{\mu^\star,\nu_0}\left(V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-\right) + 2\rho^\top\beta^{\mu^\star,\nu_0} \\
&\leq \gamma^2\rho^\top\left(P^{\mu^\star,\nu_0}\right)^2\left(V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-\right) + 2\rho^\top\beta^{\mu^\star,\nu_0} + 2\gamma\rho^\top P^{\mu^\star,\nu_0}\beta^{\mu^\star,\nu_0} \\
&\leq \cdots \leq \gamma^n\rho^\top\left(P^{\mu^\star,\nu_0}\right)^n\left(V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-\right) + 2\rho^\top\left[\sum_{i=0}^{n-1}\gamma^i\left(P^{\mu^\star,\nu_0}\right)^i\right]\beta^{\mu^\star,\nu_0}
\end{aligned}$$

holds for all positive integers n. Letting $n \to \infty$ and recalling that the vector $d^{\mu^\star,\nu_0} := [d^{\mu^\star,\nu_0}(s;\rho)]_{s\in\mathcal{S}}$ obeys (see (2.1)))

$$d^{\mu^\star,\nu_0} = (1-\gamma)\rho^\top\sum_{i=0}^{\infty}\gamma^i(P^{\mu^\star,\nu_0})^i = (1-\gamma)\rho^\top\left(I - \gamma P^{\mu^\star,\nu_0}\right)^{-1}, \tag{B.20}$$

we arrive at

$$\begin{aligned}
\rho^\top\left(V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-\right) &\leq \left\{\lim_{n\to\infty}\gamma^n\rho^\top\left(P^{\mu^\star,\nu_0}\right)^n\left(V^{\mu^\star,\nu_0} - V_{\mathsf{pe}}^-\right)\right\} + \frac{2}{1-\gamma}\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0} \\
&= \frac{2}{1-\gamma}\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0},
\end{aligned} \tag{B.21}$$

where the last line makes use of the fact that $\|\rho^\top\left(P^{\mu^\star,\nu_0}\right)^n\|_1 = 1$ for any $n \geq 1$ and hence $\gamma^n\rho^\top\left(P^{\mu^\star,\nu_0}\right)^n \to 0$ as $n \to \infty$ when $\gamma < 1$.

In order to further control (B.21), we resort to the following lemma for bounding $\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0}$, whose proof can be found in Appendix C.6.

Lemma 5. *There exists some large enough universal constant $c_6 > 0$ such that*

$$\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0} \leq c_6\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta} + c_6\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)}\log\frac{N}{(1-\gamma)\delta}}.$$

with probability exceeding $1-\delta$.

To finish up, taking (B.17), (B.21) and Lemma 5 together gives

$$V^{\star}(\rho) - V^{\widehat{\mu},\star}(\rho) = \rho^{\top}\left(V^{\star} - V^{\widehat{\mu},\star}\right) \leq \rho^{\top}\left(V^{\mu^{\star},\nu_0} - V^{-}_{\mathsf{pe}}\right) \leq \frac{2}{1-\gamma}\left(d^{\mu^{\star},\nu_0}\right)^{\top}\beta^{\mu^{\star},\nu_0}$$
$$\leq 2c_6\sqrt{\frac{C^{\star}_{\mathsf{clipped}}S\left(A+B\right)}{N\left(1-\gamma\right)^3}\log\frac{N}{\left(1-\gamma\right)\delta}} + \frac{2c_6 C^{\star}_{\mathsf{clipped}}S\left(A+B\right)}{\left(1-\gamma\right)^2 N}\log\frac{N}{\left(1-\gamma\right)\delta}.$$

This has completed the proof for the claim (B.1a). The proof for the other claim (B.1b) follows from an almost identical argument, and is hence omitted.

C Auxiliary lemmas for Theorem 1

C.1 Proof for the preliminary facts in Appendix B.1

We start by proving (B.3). For any $s \in \mathcal{S}$, suppose that a NE of the matrix game $Q_1(s,\cdot,\cdot)$ is given by $\big(\mu_1(s),\nu_1(s)\big)$, and that a NE of the matrix game $Q_2(s,\cdot,\cdot)$ is $\big(\mu_2(s),\nu_2(s)\big)$. Then we can derive

$$\begin{aligned}
V_1(s) - V_2(s) &= \mathop{\mathbb{E}}_{a\sim\mu_1(s),b\sim\nu_1(s)}\big[Q_1(s,a,b)\big] - \mathop{\mathbb{E}}_{a\sim\mu_2(s),b\sim\nu_2(s)}\big[Q_2(s,a,b)\big] \\
&\leq \mathop{\mathbb{E}}_{a\sim\mu_1(s),b\sim\nu_2(s)}\big[Q_1(s,a,b)\big] - \mathop{\mathbb{E}}_{a\sim\mu_1(s),b\sim\nu_2(s)}\big[Q_2(s,a,b)\big] \\
&= \mathop{\mathbb{E}}_{a\sim\mu_1(s),b\sim\nu_2(s)}\big[Q_1(s,a,b) - Q_2(s,a,b)\big] \leq \left\|Q_1 - Q_2\right\|_{\infty},
\end{aligned}$$

and similarly,

$$\begin{aligned}
V_1(s) - V_2(s) &\geq \mathop{\mathbb{E}}_{a\sim\mu_2(s),b\sim\nu_1(s)}\big[Q_1(s,a,b)\big] - \mathop{\mathbb{E}}_{a\sim\mu_2(s),b\sim\nu_1(s)}\big[Q_2(s,a,b)\big] \\
&= \mathop{\mathbb{E}}_{a\sim\mu_2(s),b\sim\nu_1(s)}\big[Q_1(s,a,b) - Q_2(s,a,b)\big] \geq -\left\|Q_1 - Q_2\right\|_{\infty}.
\end{aligned}$$

Taking the above two inequalities collectively yields

$$\left|V_1(s) - V_2(s)\right| \leq \left\|Q_1 - Q_2\right\|_{\infty}$$

for any $s \in \mathcal{S}$, allowing us to conclude that

$$\left\|V_1 - V_2\right\|_{\infty} \leq \left\|Q_1 - Q_2\right\|_{\infty}.$$

Next, we turn attention to proving (B.4). Towards this, we observe from (A.1) that

$$\begin{aligned}
\left|\mathsf{Var}_{P_{s,a,b}}(V_1) - \mathsf{Var}_{P_{s,a,b}}(V_2)\right| &= \left|P_{s,a,b}(V_1\circ V_1) - (P_{s,a,b}V_1)^2 - P_{s,a,b}(V_2\circ V_2) + (P_{s,a,b}V_2)^2\right| \\
&\leq \left|P_{s,a,b}(V_1\circ V_1 - V_2\circ V_2)\right| + \left|(P_{s,a,b}V_1)^2 - (P_{s,a,b}V_2)^2\right| \\
&\leq \left|P_{s,a,b}[(V_1+V_2)\circ(V_1-V_2)]\right| + \left|P_{s,a,b}(V_1-V_2)\right|\cdot\left|P_{s,a,b}(V_1+V_2)\right| \\
&\leq \frac{2}{1-\gamma}\left|P_{s,a,b}(V_1-V_2)\right| + \frac{2}{1-\gamma}\left|P_{s,a,b}(V_1-V_2)\right| \\
&\leq \frac{4}{1-\gamma}\left\|P_{s,a,b}\right\|_1\left\|V_1-V_2\right\|_{\infty} = \frac{4}{1-\gamma}\left\|V_1-V_2\right\|_{\infty},
\end{aligned} \tag{C.1}$$

whereas the validity of the last line is guaranteed since $P_{s,a,b} \in \Delta(\mathcal{S})$.

Finally, we present the proof of (B.5). Consider first the case with

$$\max\left\{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1), \mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1)\right\} < \frac{4C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)^2 N(s,a,b)}.$$

In this case, the penalty terms reduce to (cf. (3.5))

$$\beta(s,a,b;V_1) = \beta(s,a,b;V_2) = \min\left\{\frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)}, \frac{1}{1-\gamma}\right\},$$

and as a result,

$$|\beta(s,a,b;V_1) - \beta(s,a,b;V_2)| = 0 \le 2\|V_1 - V_2\|_\infty.$$

In contrast, in the other case where

$$\max\left\{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1), \mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1)\right\} \ge \frac{4C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a,b)}, \tag{C.2}$$

we can obtain

$$\begin{aligned}|\beta(s,a,b;V_1) - \beta(s,a,b;V_2)| &\le \sqrt{\frac{C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a,b)}}\left(\sqrt{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1)} - \sqrt{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_2)}\right)\\ &= \sqrt{\frac{C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a,b)}}\frac{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1) - \mathsf{Var}_{\widehat{P}_{s,a,b}}(V_2)}{\sqrt{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1)} + \sqrt{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_2)}}\\ &\le \frac{1-\gamma}{2}\left[\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_1) - \mathsf{Var}_{\widehat{P}_{s,a,b}}(V_2)\right]\\ &\le 2\|V_1 - V_2\|_\infty.\end{aligned}$$

Here, the penultimate relation follows from (C.2), while the last line takes advantage of (C.1). The two cases taken together establish (B.5).

C.2 Proof of Lemma 1

For convenience of presentation, let us define

$$\widetilde{\mathcal{T}}^{+}_{\mathsf{pe}}(Q)(s,a,b) \coloneqq \widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V + \beta(s,a,b;V), \tag{C.3a}$$

$$\widetilde{\mathcal{T}}^{-}_{\mathsf{pe}}(Q)(s,a,b) \coloneqq \widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V - \beta(s,a,b;V), \tag{C.3b}$$

where V is the value function associated to Q (see (3.3) for the definition). It is readily seen that

$$\widehat{\mathcal{T}}^{+}_{\mathsf{pe}}(Q)(s,a,b) = \min\left\{\widetilde{\mathcal{T}}^{+}_{\mathsf{pe}}(Q)(s,a,b), \frac{1}{1-\gamma}\right\}, \tag{C.4a}$$

$$\widehat{\mathcal{T}}^{-}_{\mathsf{pe}}(Q)(s,a,b) = \max\left\{\widetilde{\mathcal{T}}^{-}_{\mathsf{pe}}(Q)(s,a,b), 0\right\}. \tag{C.4b}$$

Property 1: monotonicity. In view of (C.4), it suffices to show that $\widetilde{\mathcal{T}}^{+}_{\mathsf{pe}}$ and $\widetilde{\mathcal{T}}^{-}_{\mathsf{pe}}$ are monotone. For any Q_1 and Q_2 satisfying $0 \le Q_1 \le Q_2 \le \frac{1}{1-\gamma}1$, we denote by V_1 (resp. V_2) the value function corresponds to Q_1 (resp. Q_2); see (3.3) for the precise definition. It is straightforward to check that for all $s \in \mathcal{S}$, one has $0 \le V_1, V_2 \le \frac{1}{1-\gamma}1$ and

$$\begin{aligned}V_1(s) &= \max_{\mu(s)\in\Delta(\mathcal{A})}\min_{\nu(s)\in\Delta(\mathcal{B})}\mathop{\mathbb{E}}_{a\sim\mu(s),b\sim\nu(s)}\left[Q_1(s,a,b)\right]\\ &\ge \max_{\mu(s)\in\Delta(\mathcal{A})}\min_{\nu(s)\in\Delta(\mathcal{B})}\mathop{\mathbb{E}}_{a\sim\mu(s),b\sim\nu(s)}\left[Q_2(s,a,b)\right] = V_2(s).\end{aligned}$$

For any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, define two functions $f^+, f^- : \mathbb{R}^S \to \mathbb{R}$ as

$$f^+(V) \coloneqq \widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V + \beta(s,a,b;V)$$

$$= \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V + \min\left\{ \max\left\{ \sqrt{\frac{C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a,b)} \mathsf{Var}_{\widehat{P}_{s,a,b}}(V)}, \frac{2C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma) N(s,a,b)} \right\}, \frac{1}{1-\gamma} \right\} + \frac{4}{N};$$

$$f^{-}(V) := \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V - \beta(s,a,b;V)$$

$$= \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V - \min\left\{ \max\left\{ \sqrt{\frac{C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a,b)} \mathsf{Var}_{\widehat{P}_{s,a,b}}(V)}, \frac{2C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma) N(s,a,b)} \right\}, \frac{1}{1-\gamma} \right\} - \frac{4}{N}.$$

Recognizing that $\widetilde{\mathcal{T}}_{\mathsf{pe}}^{-}(Q)(s,a,b) = f^{-}(V)$ and $\widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q)(s,a,b) = f^{+}(V)$, we only need to prove that f^{+} and f^{-} are monotone in V, namely, for any $V_1 \geq V_2$, we have $f^{+}(V_1) \geq f^{+}(V_2)$ and $f^{-}(V_1) \geq f^{-}(V_2)$. Since both f^{+} and f^{-} are continuous, it suffices to check that $\nabla f^{+}(V) \geq 0$ and $\nabla f^{-}(V) \geq 0$ hold almost everywhere.

- We first consider the case where

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}(V) < \frac{4C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a,b)}.$$

  In this case, it is seen that

$$f^{+}(V) = \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V + \min\left\{ \frac{2C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma) N(s,a,b)}, \frac{1}{1-\gamma} \right\} + \frac{4}{N},$$

$$f^{-}(V) = \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V - \min\left\{ \frac{2C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma) N(s,a,b)}, \frac{1}{1-\gamma} \right\} - \frac{4}{N},$$

  which immediately gives

$$\nabla f^{+}(V) = \nabla f^{-}(V) = \gamma \widehat{P}_{s,a,b}^{\top} \geq 0.$$

- Next, we look at the case where

$$\frac{4C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a,b)} < \mathsf{Var}_{\widehat{P}_{s,a,b}}(V) < \frac{N(s,a,b)}{C_{\mathsf{b}} (1-\gamma)^2 \log \frac{N}{(1-\gamma)\delta}}. \tag{C.5}$$

  In this case, we can easily derive

$$f^{+}(V) = \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V + \sqrt{\frac{C_{\mathsf{b}} \iota}{N(s,a,b)} \mathsf{Var}_{\widehat{P}_{s,a,b}}(V)} + \frac{5}{N}$$

$$= \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V + \sqrt{\frac{C_{\mathsf{b}} \iota}{N(s,a,b)} \left[\widehat{P}_{s,a,b}(V \circ V) - \left(\widehat{P}_{s,a,b} V \right)^2 \right]} + \frac{5}{N},$$

  and similarly,

$$f^{-}(V) = \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V - \sqrt{\frac{C_{\mathsf{b}} \iota}{N(s,a,b)} \left[\widehat{P}_{s,a,b}(V \circ V) - \left(\widehat{P}_{s,a,b} V \right)^2 \right]} - \frac{5}{N}.$$

  Recognizing that (C.5) guarantees

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}(V) = \widehat{P}_{s,a,b}(V \circ V) - \left(\widehat{P}_{s,a,b} V \right)^2 > 0,$$

  we can invoke a little linear algebra to show that

$$\nabla f^{+}(V) = \gamma \widehat{P}_{s,a,b}^{\top} + \sqrt{\frac{C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a,b)}} \frac{\widehat{P}_{s,a,b}^{\top} \circ V - \left(\widehat{P}_{s,a,b} V \right) \widehat{P}_{s,a,b}^{\top}}{\sqrt{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V)}}$$

$$\overset{\text{(i)}}{\geq} \gamma \widehat{P}_{s,a,b}^{\top} - \frac{1}{2(1-\gamma)}\big(\widehat{P}_{s,a,b}V\big)\widehat{P}_{s,a,b}^{\top} \overset{\text{(ii)}}{\geq} \gamma \widehat{P}_{s,a,b}^{\top} - \frac{1}{2}\widehat{P}_{s,a,b}^{\top} \overset{\text{(iii)}}{\geq} 0,$$

where (i) follows from (C.5), (ii) holds since $\big|\widehat{P}_{s,a,b}V\big| \leq \|V\|_\infty \leq \frac{1}{1-\gamma}$, and (iii) is valid as long as $\gamma \geq 1/2$. Similarly, we can also deduce that

$$\nabla f^{-}(V) = \gamma \widehat{P}_{s,a,b}^{\top} - \sqrt{\frac{C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a,b)}} \frac{\widehat{P}_{s,a,b}^{\top} \circ V - \big(\widehat{P}_{s,a,b}V\big)\widehat{P}_{s,a,b}^{\top}}{\sqrt{\mathsf{Var}_{\widehat{P}_{s,a,b}}(V)}}$$

$$\overset{\text{(iv)}}{\geq} \gamma \widehat{P}_{s,a,b}^{\top} - \frac{1}{2(1-\gamma)}\widehat{P}_{s,a,b}^{\top} \circ V \overset{\text{(v)}}{\geq} \gamma \widehat{P}_{s,a,b}^{\top} - \frac{1}{2}\widehat{P}_{s,a,b}^{\top} \overset{\text{(vi)}}{\geq} 0,$$

where (iv) arises from (C.5), (v) holds due to $\|V\|_\infty \leq \frac{1}{1-\gamma}$, and (vi) is guaranteed to hold when $\gamma \geq 1/2$.

- Lastly, we are left with the case with

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}(V) > \frac{N(s,a,b)}{C_{\mathsf{b}}(1-\gamma)^2 \log \frac{N}{(1-\gamma)\delta}},$$

which necessarily satisfies

$$f^{+}(V) = \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b}V + \frac{1}{1-\gamma} + \frac{5}{N},$$
$$f^{-}(V) = \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b}V - \frac{1}{1-\gamma} - \frac{5}{N}.$$

This immediately gives

$$\nabla f^{+}(V) = \nabla f^{-}(V) = \gamma \widehat{P}_{s,a,b}^{\top} \geq 0.$$

Thus far, we have verified that $\nabla f^{+}(V) \geq 0$ and $\nabla f^{-}(V) \geq 0$ hold almost everywhere, except when V is in a Legesgue zero-measure set

$$\left\{ V \in \mathbb{R}^S : \mathsf{Var}_{\widehat{P}_{s,a,b}}(V) = \frac{N(s,a,b)}{C_{\mathsf{b}}(1-\gamma)^2 \log \frac{N}{(1-\gamma)\delta}} \quad \text{or} \quad \mathsf{Var}_{\widehat{P}_{s,a,b}}(V) = \frac{4C_{\mathsf{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a,b)} \right\}.$$

This in turn completes the proof of the claimed monotonicity property.

Property 2: γ-contraction. We shall only prove the γ-contraction property for $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}$; the proof w.r.t. $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ follows from an analogous argument and is hence omitted. In view of the identities (C.4), it is straightforward to check that

$$\left\| \widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2) \right\|_\infty \leq \left\| \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2) \right\|_\infty$$

for any $Q_1, Q_2 : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \big[0, \frac{1}{1-\gamma}\big]$. Therefore, it suffices to show that

$$\left\| \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2) \right\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty . \tag{C.6}$$

Recalling that we have already shown above that $\widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}$ is monotone, namely,

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) \geq \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2), \qquad \forall\, Q_1 \geq Q_2, \tag{C.7}$$

we can immediately apply it and the triangle inequality to obtain

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2) \leq \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}\big(Q_2 + \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2). \tag{C.8}$$

Letting V_2 be the value function corresponding to Q_2, we can straightforwardly check that the value function V^{upper} associated with the vector $Q^{\mathsf{upper}} := Q_2 + \|Q_1 - Q_2\|_\infty 1$ (cf. (3.3)) is given by

$$\begin{aligned} V^{\mathsf{upper}}(s) &= \max_{\mu(s)\in\Delta(\mathcal{A})} \min_{\nu(s)\in\Delta(\mathcal{B})} \mathop{\mathbb{E}}_{a\sim\mu(s),b\sim\nu(s)} \left[Q_2(s,a,b) + \|Q_1 - Q_2\|_\infty 1\right] \\ &= \|Q_1 - Q_2\|_\infty + \max_{\mu(s)\in\Delta(\mathcal{A})} \min_{\nu(s)\in\Delta(\mathcal{B})} \mathop{\mathbb{E}}_{a\sim\mu(s),b\sim\nu(s)} \left[Q_2(s,a,b)\right] \\ &= V_2(s) + \|Q_1 - Q_2\|_\infty . \end{aligned}$$

Additionally, we can further rewrite the penalty term as

$$\begin{aligned} \beta(s,a,b;V_2) &= \min\left\{\max\left\{\sqrt{\frac{C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a,b)}\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_2)}, \frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)}\right\}, \frac{1}{1-\gamma}\right\} + \frac{5}{N} \\ &= \min\left\{\max\left\{\sqrt{\frac{C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a,b)}\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_2 + \|Q_1 - Q_2\|_\infty 1)}, \frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)}\right\}, \frac{1}{1-\gamma}\right\} + \frac{5}{N} \\ &= \beta(s,a,b;V_2 + \|Q_1 - Q_2\|_\infty 1), \end{aligned}$$

which combined with (C.3a) gives

$$\left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2 + \|Q_1 - Q_2\|_\infty 1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2)\right\|_\infty = \gamma\left\|\widehat{P}_{s,a,b}(\|Q_1 - Q_2\|_\infty 1)\right\|_\infty = \gamma\|Q_1 - Q_2\|_\infty . \tag{C.9}$$

Taking (C.8) and (C.9) collectively yields

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2) \le \gamma\|Q_1 - Q_2\|_\infty 1. \tag{C.10}$$

Similarly, we can also show that

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2) \ge \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2 - \|Q_1 - Q_2\|_\infty 1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q_2) \ge -\gamma\|Q_1 - Q_2\|_\infty 1.$$

Combine (C.9) and (C.10) to establish (C.6).

Property 3: existence and uniqueness of the fixed point. For any $0 \le Q \le \frac{1}{1-\gamma}1$, we know that its associated value function V (cf. (3.3)) satisfies $0 \le V(s) \le \frac{1}{1-\gamma}$ for any $s \in \mathcal{S}$. Therefore, it is seen that

$$\begin{aligned} \widetilde{\mathcal{T}}_{\mathsf{pe}}^{+}(Q)(s,a,b) &\ge \widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V \ge 0, \\ \widetilde{\mathcal{T}}_{\mathsf{pe}}^{-}(Q)(s,a,b) &\le \widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V \le 1 + \frac{\gamma}{1-\gamma} \le \frac{1}{1-\gamma}. \end{aligned}$$

Combining the above relations with (C.4) gives

$$0 \le \widehat{\mathcal{T}}_{\mathsf{pe}}^{+}(Q)(s,a,b) \le \frac{1}{1-\gamma} \qquad \text{and} \qquad 0 \le \widehat{\mathcal{T}}_{\mathsf{pe}}^{-}(Q)(s,a,b) \le \frac{1}{1-\gamma}.$$

This together with the γ-contraction property indicates that $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+}$ and $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ are both contraction mappings on the complete metric space $\left([0, \frac{1}{1-\gamma}]^{SAB}, \|\cdot\|_\infty\right)$. In view of the Banach fixed point theorem (see, e.g., Ciesielski (2007, Theorem 3.1)), both operators admit unique fixed points, which we shall denote by $Q_{\mathsf{pe}}^{+\star}$ and $Q_{\mathsf{pe}}^{-\star}$ respectively. Clearly, it holds that $0 \le Q_{\mathsf{pe}}^{+\star} \le \frac{1}{1-\gamma}1$ and $0 \le Q_{\mathsf{pe}}^{-\star} \le \frac{1}{1-\gamma}1$.

C.3 Proof of Lemma 2

The proof of Lemma 2, which is based upon the leave-one-out analysis framework (Agarwal et al., 2020; Li et al., 2020), follows from similar arguments as that of Li et al. (2022c, Lemma 8), the latter of which can be viewed as a single-agent version of the current lemma. Given that similarity between the proofs of these two lemmas, we shall only emphasize here the difference between the two proofs for the sake of brevity; the interested reader is referred to Li et al. (2022c, Appendix B.4) for details.

We first construct a set of auxiliary Markov games. For any $s \in \mathcal{S}$ and any $u > 0$, we define a Markov game $\mathcal{MG}^{s,u} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P^{s,u}, r^{s,u}, \gamma)$ as follows: the transition kernel $P^{s,u}$ is defined as

$$\begin{aligned} P^{s,u}(\widetilde{s}\,|\,s,a,b) &= \mathbb{1}\{\widetilde{s} = s\}, && \text{for all } (\widetilde{s},a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \\ P^{s,u}(\cdot\,|\,s',a,b) &= \widehat{P}(\cdot\,|\,s',a,b), && \text{for all } (s',a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \text{ and } s' \neq s; \end{aligned}$$

and the reward function $r^{s,u}$ is defined as

$$\begin{aligned} r^{s,u}(s,a,b) &= u, && \text{for all } (a,b) \in \mathcal{A}\times\mathcal{B}, \\ r^{s,u}(s',a,b) &= \widehat{r}(s',a,b), && \text{for all } (s',a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \text{ and } s' \neq s. \end{aligned}$$

Define the corresponding pessimistic operators $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-,s,u}, \widehat{\mathcal{T}}_{\mathsf{pe}}^{+,s,u} : \mathbb{R}^{SAB} \to \mathbb{R}^{SAB}$ for the max- and min-player, respectively: for any $Q: \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \mathbb{R}$ and any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, let

$$\begin{aligned} \widehat{\mathcal{T}}_{\mathsf{pe}}^{-,s,u}(Q)(s,a,b) &:= \max\left\{ r^{s,u}(s,a,b) + \gamma P^{s,u}_{s,a,b} V - \beta^{s,u}(s,a,b;V), 0 \right\}, \\ \widehat{\mathcal{T}}_{\mathsf{pe}}^{+,s,u}(Q)(s,a,b) &:= \min\left\{ r^{s,u}(s,a,b) + \gamma P^{s,u}_{s,a,b} V + \beta^{s,u}(s,a,b;V), \frac{1}{1-\gamma} \right\}, \end{aligned}$$

where $V: \mathcal{S} \to \mathbb{R}$ is the value function associated with Q (see (3.3) for the definition), and

$$\beta^{s,u}(s,a,b;V) = \min\left\{ \max\left\{ \sqrt{\frac{C_{\mathsf{b}} \log\frac{N}{(1-\gamma)\delta}}{N(s,a,b)} \mathsf{Var}_{P^{s,u}_{s,a,b}}(V)}, \frac{2C_{\mathsf{b}} \log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)} \right\}, \frac{1}{1-\gamma} \right\} + \frac{4}{N}.$$

With the above construction in place, we can follow the similar analysis as Step 2 in Li et al. (2022c, Appendix B.4) to show that: by taking

$$\begin{aligned} u^{-\star} &= (1-\gamma) V_{\mathsf{pe}}^{-\star}(s) + \min\left\{ \frac{2C_{\mathsf{b}} \log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)}, \frac{1}{1-\gamma} \right\} + \frac{4}{N}, \\ u^{+\star} &= (1-\gamma) V_{\mathsf{pe}}^{+\star}(s) - \min\left\{ \frac{2C_{\mathsf{b}} \log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a,b)}, \frac{1}{1-\gamma} \right\} - \frac{4}{N}, \end{aligned}$$

there exists a fixed point $Q^{-\star}_{s,u^{-\star}}$ (resp. $Q^{+\star}_{s,u^{+\star}}$) of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-,s,u}$ (resp. $\widehat{\mathcal{T}}_{\mathsf{pe}}^{+,s,u}$) whose corresponding value function $V^{-\star}_{s,u^{-\star}}$ (resp. $V^{+\star}_{s,u^{+\star}}$) coincides with $V_{\mathsf{pe}}^{-\star}$ (resp. $V_{\mathsf{pe}}^{+\star}$). Then, repeating the same ε-covering argument as in Step 3-5 of Li et al. (2022c, Appendix B.4) allows us to complete the proof.

C.4 Proof of Lemma 3

To begin with, we would like to justify that for any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, it holds that

$$Q^{\widehat{\mu},\star}(s,a,b) \geq Q_{\mathsf{pe}}^{-}(s,a,b).$$

Recall from Lemma 1 that $Q_{\mathsf{pe}}^{-\star}$ is the fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{-}$ and hence

$$Q_{\mathsf{pe}}^{-\star}(s,a,b) = \left(\widehat{\mathcal{T}}_{\mathsf{pe}}^{-} Q_{\mathsf{pe}}^{-\star}\right)(s,a,b) = \max\left\{ \widehat{r}(s,a,b) + \gamma \widehat{P}_{s,a,b} V_{\mathsf{pe}}^{-\star} - \beta\left(s,a,b; V_{\mathsf{pe}}^{-\star}\right), 0 \right\}. \tag{C.11}$$

If $Q_{\mathsf{pe}}^{-\star}(s,a,b) = 0$, then we can utilize the facts that $Q_{\mathsf{pe}}^{-} \leq Q_{\mathsf{pe}}^{-\star}$ (as shown in (B.6)) and $Q^{\widehat{\mu},\star}(s,a,b) \geq 0$ to ensure that

$$Q_{\mathsf{pe}}^{-}(s,a,b) \leq Q_{\mathsf{pe}}^{-\star}(s,a,b) = 0 \leq Q^{\widehat{\mu},\star}(s,a,b).$$

Therefore, it suffices to look at the case when $Q_{\mathsf{pe}}^{-\star}(s,a,b) > 0$ in the sequel.

Note that the condition $Q_{\mathsf{pe}}^{-\star}(s,a,b) > 0$ taken collectively with (C.11) gives

$$Q_{\mathsf{pe}}^{-\star}\left(s,a,b\right) = \widehat{r}\left(s,a,b\right) + \gamma \widehat{P}_{s,a,b} V_{\mathsf{pe}}^{-\star} - \beta\left(s,a,b; V_{\mathsf{pe}}^{-\star}\right). \tag{C.12}$$

In addition, in this case we must have $N(s,a,b) > 0$: otherwise we have

$$\beta\left(s,a,b; V_{\mathsf{pe}}^{-\star}\right) = \frac{1}{1-\gamma} + \frac{3}{N} > \frac{1}{1-\gamma},$$

which taken collectively with (C.12) leads to the following contradiction

$$Q_{\mathsf{pe}}^{-\star}\left(s,a,b\right) = \widehat{r}\left(s,a,b\right) + \gamma \widehat{P}_{s,a,b} V_{\mathsf{pe}}^{-\star} - \beta\left(s,a,b; V_{\mathsf{pe}}^{-\star}\right) < \frac{\gamma}{1-\gamma} - \frac{1}{1-\gamma} < 0.$$

With the condition $N(s,a,b) > 0$ in mind, we can proceed to bound

$$\begin{aligned}
Q_{\mathsf{pe}}^{-}\left(s,a,b\right) &\overset{\text{(i)}}{\leq} Q_{\mathsf{pe}}^{-\star}\left(s,a,b\right) \overset{\text{(ii)}}{=} \widehat{r}\left(s,a,b\right) + \gamma \widehat{P}_{s,a,b} V_{\mathsf{pe}}^{-\star} - \beta\left(s,a,b; V_{\mathsf{pe}}^{-\star}\right) \\
&\leq \widehat{r}\left(s,a,b\right) + \gamma \widehat{P}_{s,a,b} V_{\mathsf{pe}}^{-} - \beta\left(s,a,b; V_{\mathsf{pe}}^{-\star}\right) + \gamma \left\| V_{\mathsf{pe}}^{-} - V_{\mathsf{pe}}^{-\star} \right\|_{\infty} \\
&\overset{\text{(iii)}}{\leq} \widehat{r}\left(s,a,b\right) + \gamma P_{s,a,b} V_{\mathsf{pe}}^{-} - \beta\left(s,a,b; V_{\mathsf{pe}}^{-\star}\right) + \gamma \big(\widehat{P}_{s,a,b} - P_{s,a,b}\big) V_{\mathsf{pe}}^{-} + \frac{1}{N} \\
&\overset{\text{(iv)}}{\leq} \widehat{r}\left(s,a,b\right) + \gamma P_{s,a,b} V_{\mathsf{pe}}^{-} - \beta\left(s,a,b; V_{\mathsf{pe}}^{-}\right) + \gamma \big(\widehat{P}_{s,a,b} - P_{s,a,b}\big) V_{\mathsf{pe}}^{-} + \frac{3}{N} \\
&\overset{\text{(v)}}{\leq} r\left(s,a,b\right) + \gamma P_{s,a,b} V_{\mathsf{pe}}^{-}.
\end{aligned} \tag{C.13}$$

Here (i) follows from (B.6); (ii) holds due to (C.12); (iii) invokes (B.7); (iv) makes use of (B.5) and (B.7); and (v) comes from (B.11) and the fact that $\widehat{r}(s,a,b) = r(s,a,b)$ when $N(s,a,b) \geq 1$. As a result, we obtain

$$\begin{aligned}
Q^{\widehat{\mu},\star}\left(s,a,b\right) - Q_{\mathsf{pe}}^{-}\left(s,a,b\right) &= r\left(s,a,b\right) + \gamma P_{s,a,b} V^{\widehat{\mu},\star} - Q_{\mathsf{pe}}^{-}\left(s,a,b\right) \\
&\geq \gamma P_{s,a,b} \big(V^{\widehat{\mu},\star} - V_{\mathsf{pe}}^{-}\big),
\end{aligned} \tag{C.14}$$

where the first identity follows from the Bellman equation, and the last relation holds due to (C.13). Let us take

$$(s_0, a_0, b_0) = \underset{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}}{\arg\min} \left\{ Q^{\widehat{\mu},\star}\left(s,a,b\right) - Q_{\mathsf{pe}}^{-}\left(s,a,b\right) \right\}, \tag{C.15}$$

and use the notation ν_{br} to denote the best-response policy of the min-player when the max-player adopts policy $\widehat{\mu}$, i.e., $\nu_{\mathsf{br}} := \arg\min_{\nu} V^{\widehat{\mu},\nu}$ (whose existence is guaranteed by the existence of optimal policy in MDP). Consequently, we arrive at

$$\begin{aligned}
Q^{\widehat{\mu},\star}\left(s_0,a_0,b_0\right) - Q_{\mathsf{pe}}^{-}\left(s_0,a_0,b_0\right) &\overset{\text{(i)}}{\geq} \gamma P_{s_0,a_0,b_0} \big(V^{\widehat{\mu},\star} - V_{\mathsf{pe}}^{-}\big) \geq \gamma \min_{s\in\mathcal{S}} \left[V^{\widehat{\mu},\star}\left(s\right) - V_{\mathsf{pe}}^{-}\left(s\right) \right] \\
&= \gamma \min_{s\in\mathcal{S}} \left\{ \mathop{\mathbb{E}}_{a\sim\widehat{\mu}(s), b\sim\nu_{\mathsf{br}}(s)} \left[Q^{\widehat{\mu},\star}\left(s,a,b\right) \right] - \mathop{\mathbb{E}}_{a\sim\mu_T^{-}(s), b\sim\nu_T^{-}(s)} \left[Q_{\mathsf{pe}}^{-}\left(s,a,b\right) \right] \right\} \\
&\overset{\text{(ii)}}{\geq} \gamma \min_{s\in\mathcal{S}} \mathop{\mathbb{E}}_{a\sim\widehat{\mu}(s), b\sim\nu_{\mathsf{br}}(s)} \left[Q^{\widehat{\mu},\star}\left(s,a,b\right) - Q_{\mathsf{pe}}^{-}\left(s,a,b\right) \right] \\
&\geq \gamma \min_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} \left[Q^{\widehat{\mu},\star}\left(s,a,b\right) - Q_{\mathsf{pe}}^{-}\left(s,a,b\right) \right] \\
&\overset{\text{(iii)}}{=} \gamma \left[Q^{\widehat{\mu},\star}\left(s_0,a_0,b_0\right) - Q_{\mathsf{pe}}^{-}\left(s_0,a_0,b_0\right) \right].
\end{aligned} \tag{C.16}$$

Here (i) relies on (C.14); (ii) holds true since $(\mu_T^{-}(s), \nu_T^{-}(s))$ is the Nash equilibrium of $Q_{\mathsf{pe}}^{-}(s,\cdot,\cdot)$ and recall that $\widehat{\mu} = \mu_T^{-}$ and hence

$$\mathop{\mathbb{E}}_{a\sim\mu_T^{-}(s), b\sim\nu_T^{-}(s)} \left[Q_{\mathsf{pe}}^{-}\left(s,a,b\right) \right] \leq \mathop{\mathbb{E}}_{a\sim\widehat{\mu}(s), b\sim\nu_{\mathsf{br}}(s)} \left[Q^{\widehat{\mu},\star}\left(s,a,b\right) - Q_{\mathsf{pe}}^{-}\left(s,a,b\right) \right];$$

and (iii) arises from the definition of (s_0, a_0, b_0) (cf. (C.15)). Given that $\gamma < 1$, Condition (C.16) can only hold if

$$Q^{\widehat{\mu},\star}(s_0, a_0, b_0) - Q^{-}_{\mathsf{pe}}(s_0, a_0, b_0) \ge 0,$$

Therefore, we can conclude that $Q^{\widehat{\mu},\star} \ge Q^{-}_{\mathsf{pe}}$.

Similarly, we can also show that $Q^{+}_{\mathsf{pe}} \ge Q^{\star,\widehat{\nu}}$ via the same argument, which we omit here for the sake of conciseness. The claimed properties regarding the V-function are therefore immediately consequences from $Q^{\widehat{\mu},\star} \ge Q^{-}_{\mathsf{pe}}$ and $Q^{+}_{\mathsf{pe}} \ge Q^{\star,\widehat{\nu}}$.

C.5 Proof of Lemma 4

It is first observed from (B.12) and (B.13) that

$$V^{-}_{\mathsf{pe}}(s) \ge \mathop{\mathbb{E}}_{a\sim\mu^\star(s), b\sim\nu_0(s)}\left[Q^{-}_{\mathsf{pe}}(s,a,b)\right] \ge \mathop{\mathbb{E}}_{a\sim\mu^\star(s), b\sim\nu_0(s)}\left[Q^{-\star}_{\mathsf{pe}}(s,a,b)\right] - \frac{1}{N}, \tag{C.17}$$

where the last inequality comes from (B.7).

- For any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ obeying $N(s,a,b) \ge 1$, it is seen that

$$\begin{aligned} Q^{-\star}_{\mathsf{pe}}(s,a,b) &\overset{\text{(i)}}{=} \max\left\{\widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V^{-\star}_{\mathsf{pe}} - \beta\left(s,a,b;V^{-\star}_{\mathsf{pe}}\right), 0\right\} \\ &\ge \widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V^{-\star}_{\mathsf{pe}} - \beta\left(s,a,b;V^{-\star}_{\mathsf{pe}}\right) \\ &\overset{\text{(ii)}}{\ge} \widehat{r}(s,a,b) + \gamma\widehat{P}_{s,a,b}V^{-}_{\mathsf{pe}} - \beta\left(s,a,b;V^{-}_{\mathsf{pe}}\right) - \frac{3}{N} \\ &\overset{\text{(iii)}}{\ge} r(s,a,b) + \gamma P_{s,a,b}V^{-}_{\mathsf{pe}} - 2\beta\left(s,a,b;V^{-}_{\mathsf{pe}}\right) + \frac{1}{N} \end{aligned}$$

  holds with probability exceeding $1-\delta$. Here, (i) holds true since $\widehat{Q}^{\star}_{\mathsf{pe}}$ is the fixed point of $\widehat{\mathcal{T}}^{-}_{\mathsf{pe}}$; (ii) follows from (B.7) and (B.5); and (iii) follows from (B.11) and the fact that $\widehat{r}(s,a,b) = r(s,a,b)$ when $N(s,a,b) \ge 1$.

- If instead $N(s,a,b) = 0$, then by definition we have

$$\beta\left(s,a,b;V^{-\star}_{\mathsf{pe}}\right) = \frac{1}{1-\gamma} + \frac{4}{N},$$

  and therefore,

$$Q^{-\star}_{\mathsf{pe}}(s,a,b) \ge 0 \ge r(s,a,b) + \gamma P_{s,a,b}V^{-}_{\mathsf{pe}} - 2\beta\left(s,a,b;V^{-}_{\mathsf{pe}}\right) + \frac{1}{N}$$

  holds as well.

To summarize, with probability exceeding $1-\delta$, one has

$$Q^{-\star}_{\mathsf{pe}}(s,a,b) \ge r(s,a,b) + \gamma P_{s,a,b}V^{-}_{\mathsf{pe}} - 2\beta\left(s,a,b;V^{-}_{\mathsf{pe}}\right) + \frac{1}{N}$$

holds simultaneously for all $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$. This taken collectively with (C.17) yields

$$V^{-}_{\mathsf{pe}}(s) \ge \mathop{\mathbb{E}}_{a\sim\mu^\star(s), b\sim\nu_0(s)}\left[r(s,a,b) + \gamma P_{s,a,b}V^{-}_{\mathsf{pe}} - 2\beta\left(s,a,b;V^{-}_{\mathsf{pe}}\right)\right]$$

holds for all $s \in \mathcal{S}$. This concludes the proof of (B.16).

C.6 Proof of Lemma 5

Before proceeding, we make note of a key result will be useful in the proof.

Claim 1. With probability exceeding $1-\delta$, we have

$$\max\left\{N\left(s,a,b\right),\log\frac{N}{\delta}\right\}\geq\frac{1}{2}Nd_{\mathsf{b}}\left(s,a,b\right) \tag{C.18}$$

holds for all $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$.

Proof. See Appendix C.6.1. □

With the above claim in place, we can readily make the observation that

$$\begin{aligned}
\beta\left(s,a,b;V_{\mathsf{pe}}^{-}\right) &= \min\left\{\max\left\{\sqrt{\frac{C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{N\left(s,a,b\right)}\mathsf{Var}_{\widehat{P}_{s,a,b}}\left(V_{\mathsf{pe}}^{-}\right)},\frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)N\left(s,a,b\right)}\right\},\frac{1}{1-\gamma}\right\}+\frac{4}{N}\\
&\overset{\text{(i)}}{\leq}\max\left\{\sqrt{\frac{C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{\max\left\{N\left(s,a,b\right),\log\frac{N}{\delta}\right\}}\mathsf{Var}_{\widehat{P}_{s,a,b}}\left(V_{\mathsf{pe}}^{-}\right)},\frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)\max\left\{N\left(s,a,b\right),\log\frac{N}{\delta}\right\}}\right\}+\frac{4}{N}\\
&\overset{\text{(ii)}}{\leq}\max\left\{\sqrt{\frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{Nd_{\mathsf{b}}\left(s,a,b\right)}\mathsf{Var}_{\widehat{P}_{s,a,b}}\left(V_{\mathsf{pe}}^{-}\right)},\frac{4C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)Nd_{\mathsf{b}}\left(s,a,b\right)}\right\}+\frac{4}{N}\\
&\leq c_{3}\sqrt{\frac{\log\frac{N}{(1-\gamma)\delta}}{Nd_{\mathsf{b}}\left(s,a,b\right)}\mathsf{Var}_{\widehat{P}_{s,a,b}}\left(V_{\mathsf{pe}}^{-}\right)}+\frac{c_{3}\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)Nd_{\mathsf{b}}\left(s,a,b\right)}+\frac{4}{N}
\end{aligned}$$

for some universal constant $c_3\geq\max\left\{\sqrt{2C_{\mathsf{b}}},4C_{\mathsf{b}}\right\}$. To see why (i) holds, it suffices to note that in the case where $N(s,a,b)\leq\log(N/\delta)$, it holds that

$$\frac{1}{1-\gamma}\leq\frac{2C_{\mathsf{b}}\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)\max\left\{N\left(s,a,b\right),\log\frac{N}{\delta}\right\}},$$

provided that $C_{\mathsf{b}}\geq 1$; and with regards to (ii), we take advantage of Claim 1. Consequently, we can combine this with the definition of $\beta^{\mu^\star,\nu_0}$ (cf. (B.15b)) to arrive at

$$\begin{aligned}
\left(d^{\mu^\star,\nu_0}\right)^{\top}\beta^{\mu^\star,\nu_0}\leq{}& c_3\underbrace{\sum_{s\in\mathcal{S}}d^{\mu^\star,\nu_0}\left(s;\rho\right)\mathop{\mathbb{E}}_{a\sim\mu^\star(s),b\sim\nu_0(s)}\left[\frac{\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)Nd_{\mathsf{b}}\left(s,a,b\right)}\right]}_{=:\alpha_1}+\frac{4}{N}\\
&+c_3\underbrace{\sum_{s\in\mathcal{S}}d^{\mu^\star,\nu_0}\left(s;\rho\right)\mathop{\mathbb{E}}_{a\sim\mu^\star(s),b\sim\nu_0(s)}\left[\sqrt{\frac{\log\frac{N}{(1-\gamma)\delta}}{Nd_{\mathsf{b}}\left(s,a,b\right)}\mathsf{Var}_{\widehat{P}_{s,a,b}}\left(V_{\mathsf{pe}}^{-}\right)}\right]}_{=:\alpha_2},
\end{aligned} \tag{C.19}$$

leaving us with two terms to deal with.

Bounding the first term α_1. Let us begin with the first term α_1 in (C.19). Recalling that $\nu_0:\mathcal{S}\to\mathcal{B}$ is a deterministic policy (see (B.14)), we can upper bound

$$\begin{aligned}
\alpha_1 &= \sum_{s\in\mathcal{S},a\in\mathcal{A},b\in\mathcal{B}}d^{\mu^\star,\nu_0}(s;\rho)\mu^\star(a\,|\,s)\,\mathbb{1}\left\{b=\nu_0(s)\right\}\frac{\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)Nd_{\mathsf{b}}(s,a,b)}\\
&\overset{\text{(i)}}{=}\sum_{s\in\mathcal{S},a\in\mathcal{A}}d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\frac{\log\frac{N}{(1-\gamma)\delta}}{\left(1-\gamma\right)Nd_{\mathsf{b}}\big(s,a,\nu_0(s)\big)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\frac{d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)}{d_{\mathsf{b}}\big(s,a,\nu_0(s)\big)}\\
&\overset{\text{(ii)}}{\leq} \frac{C^\star_{\mathsf{clipped}}}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\frac{d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)}{\min\Big\{d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big),\frac{1}{S(A+B)}\Big\}}\\
&\leq \frac{C^\star_{\mathsf{clipped}}SA}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta}+\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta}\sum_{s\in\mathcal{S},a\in\mathcal{A}}d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\\
&\leq 2\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta},
\end{aligned}
\tag{C.20}
$$

where (i) comes from (2.3), and (ii) follows from Assumption 2.

Bounding the second term α_2. Next, we move on to bounding the second term α_2 in (C.19), which relies on the following result.

***Claim* 2.** With probability exceeding $1-\delta$, there exists some universal constant $c_4>0$ such that

$$
\mathsf{Var}_{\widehat{P}_{s,a,b}}\big(V^-_{\mathsf{pe}}\big)\leq 2\mathsf{Var}_{P_{s,a,b}}\big(V^-_{\mathsf{pe}}\big)+\frac{c_4\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2Nd_{\mathsf{b}}(s,a,b)}
$$

holds simultaneously for all $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$.

Proof. See Appendix C.6.2. □

The bound in Claim 2 allows one to deduce that

$$
\begin{aligned}
\alpha_2 &= \sum_{s\in\mathcal{S},a\in\mathcal{A},b\in\mathcal{B}}d^{\mu^\star,\nu_0}(s;\rho)\mu^\star(a\,|\,s)\mathbb{1}\{b=\nu_0(s)\}\sqrt{\frac{\log\frac{N}{(1-\gamma)\delta}}{Nd_{\mathsf{b}}(s,a,b)}\mathsf{Var}_{\widehat{P}_{s,a,\nu_0(s)}}\big(V^-_{\mathsf{pe}}\big)}\\
&= \sum_{s\in\mathcal{S},a\in\mathcal{A}}d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\sqrt{\frac{\log\frac{N}{(1-\gamma)\delta}}{Nd_{\mathsf{b}}\big(s,a,\nu_0(s)\big)}\mathsf{Var}_{\widehat{P}_{s,a,\nu_0(s)}}\big(V^-_{\mathsf{pe}}\big)}\\
&\leq c_5\sum_{s\in\mathcal{S},a\in\mathcal{A}}d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\left[\sqrt{\frac{\log\frac{N}{(1-\gamma)\delta}}{Nd_{\mathsf{b}}\big(s,a,\nu_0(s)\big)}\mathsf{Var}_{P_{s,a,\nu_0(s)}}\big(V^-_{\mathsf{pe}}\big)}+\frac{\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)Nd_{\mathsf{b}}\big(s,a,\nu_0(s)\big)}\right]\\
&= \underbrace{c_5\sum_{s\in\mathcal{S},a\in\mathcal{A}}d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\sqrt{\frac{\log\frac{N}{(1-\gamma)\delta}}{Nd_{\mathsf{b}}\big(s,a,\nu_0(s)\big)}\mathsf{Var}_{P_{s,a,\nu_0(s)}}\big(V^-_{\mathsf{pe}}\big)}}_{=:\alpha_3}+c_5\alpha_1
\end{aligned}
\tag{C.21}
$$

for some universal constant $c_5>0$. It thus boils down to bounding α_3, which we accomplish next.

Bounding the intermediate term α_3. It is observed that

$$
\begin{aligned}
\alpha_3 &\overset{\text{(i)}}{\leq} \sum_{s\in\mathcal{S},a\in\mathcal{A}}d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\sqrt{\frac{C^\star_{\mathsf{clipped}}\log\frac{N}{(1-\gamma)\delta}}{N\min\Big\{d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big),\frac{1}{S(A+B)}\Big\}}\mathsf{Var}_{P_{s,a,\nu_0(s)}}\big(V^-_{\mathsf{pe}}\big)}\\
&\leq \sqrt{\frac{C^\star_{\mathsf{clipped}}\log\frac{N}{(1-\gamma)\delta}}{N}}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sqrt{d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\mathsf{Var}_{P_{s,a,\nu_0(s)}}\big(V^-_{\mathsf{pe}}\big)}\\
&\quad+\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)\log\frac{N}{(1-\gamma)\delta}}{N}}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\sqrt{d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)}\cdot\sqrt{d^{\mu^\star,\nu_0}\big(s,a,\nu_0(s);\rho\big)\mathsf{Var}_{P_{s,a,\nu_0(s)}}\big(V^-_{\mathsf{pe}}\big)}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{C^{\star}_{\mathsf{clipped}}}{N}\log\frac{N}{(1-\gamma)\delta}} \cdot \sqrt{SA} \cdot \sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}} d^{\mu^{\star},\nu_0}\big(s,a,\nu_0(s);\rho\big)\mathsf{Var}_{P_{s,a,\nu_0(s)}}\left(V^{-}_{\mathsf{pe}}\right)} \\
&\quad + \sqrt{\frac{C^{\star}_{\mathsf{clipped}} S\left(A+B\right)\log\frac{N}{(1-\gamma)\delta}}{N}}\left[\sum_{s\in\mathcal{S},a\in\mathcal{A}} d^{\mu^{\star},\nu_0}\big(s,a,\nu_0(s);\rho\big)\right]\sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}} d^{\mu^{\star},\nu_0}\big(s,a,\nu_0(s);\rho\big)\mathsf{Var}_{P_{s,a,\nu_0(s)}}\left(V^{-}_{\mathsf{pe}}\right)} \\
&\leq 2\sqrt{\frac{C^{\star}_{\mathsf{clipped}} S\left(A+B\right)\log\frac{N}{(1-\gamma)\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S},a\in\mathcal{A}} d^{\mu^{\star},\nu_0}\big(s,a,\nu_0(s);\rho\big)\mathsf{Var}_{P_{s,a,\nu_0(s)}}\left(V^{-}_{\mathsf{pe}}\right)} \\
&\overset{\text{(iii)}}{=} 2\sqrt{\frac{C^{\star}_{\mathsf{clipped}} S\left(A+B\right)}{N}\log\frac{N}{(1-\gamma)\delta}}\sqrt{\sum_{s\in\mathcal{S}} d^{\mu^{\star},\nu_0}\left(s;\rho\right)\mathop{\mathbb{E}}_{a\sim\mu^{\star}(s),b\sim\nu_0(s)}\left[\mathsf{Var}_{P_{s,a,b}}\left(V^{-}_{\mathsf{pe}}\right)\right]} \\
&\overset{\text{(iv)}}{\leq} 2\sqrt{\frac{C^{\star}_{\mathsf{clipped}} S\left(A+B\right)}{N}\log\frac{N}{(1-\gamma)\delta}}\sqrt{\sum_{s\in\mathcal{S}} d^{\mu^{\star},\nu_0}\left(s;\rho\right)\mathsf{Var}_{P^{\mu^{\star},\nu_0}_{s}}\left(V^{-}_{\mathsf{pe}}\right)}.
\end{aligned} \tag{C.22}
$$

Here, (i) follows from Assumption 2; (ii) holds due to the Cauchy-Schwarz inequality; (iii) is valid since $d^{\mu^{\star},\nu_0}\big(s,a,b;\rho\big) = d^{\mu^{\star},\nu_0}\big(s;\rho\big)\mu^{\star}(a\,|\,s)\,\mathbb{1}\{b=\nu_0(s)\}$; and (iv) can be justified as follows

$$
\begin{aligned}
\mathop{\mathbb{E}}_{a\sim\mu^{\star}(s),b\sim\nu_0(s)}\left[\mathsf{Var}_{P_{s,a,b}}\left(V^{-}_{\mathsf{pe}}\right)\right] &= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\mu^{\star}\left(a\,|\,s\right)\nu_0\left(b\,|\,s\right)\mathsf{Var}_{P_{s,a,b}}\left(V^{-}_{\mathsf{pe}}\right) \\
&= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\mu^{\star}\left(a\,|\,s\right)\nu_0\left(b\,|\,s\right)\left[P_{s,a,b}\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right) - \left(P_{s,a,b}V^{-}_{\mathsf{pe}}\right)^2\right] \\
&\leq \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\mu^{\star}\left(a\,|\,s\right)\nu_0\left(b\,|\,s\right)P_{s,a,b}\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right) - \left(\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\mu^{\star}\left(a\,|\,s\right)\nu_0\left(b\,|\,s\right)P_{s,a,b}V^{-}_{\mathsf{pe}}\right)^2 \\
&= P^{\mu^{\star},\nu_0}(\cdot\,|\,s)\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right) - \left(P^{\mu^{\star},\nu_0}(\cdot\,|\,s)V^{-}_{\mathsf{pe}}\right)^2 = \mathsf{Var}_{P^{\mu^{\star},\nu_0}(\cdot\,|\,s)}\left(V^{-}_{\mathsf{pe}}\right),
\end{aligned}
$$

where the penultimate line applies Jensen's inequality, and the last line relies on the definition of $P^{\mu^{\star},\nu_0}$ (see (B.15c)). Defining a vector $v = [v_s]_{s\in\mathcal{S}} \in \mathbb{R}^S$ such that $v_s := \mathsf{Var}_{P^{\mu^{\star},\nu_0}(\cdot\,|\,s)}(V^{-}_{\mathsf{pe}})$, we obtain

$$
\begin{aligned}
v &= P^{\mu^{\star},\nu_0}\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right) - \left(P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\right)\circ\left(P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\right) \\
&= P^{\mu^{\star},\nu_0}\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right) - \frac{1}{\gamma^2}V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}} + \frac{1}{\gamma^2}V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}} - \left(P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\right)\circ\left(P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\right) \\
&\overset{\text{(i)}}{\leq} P^{\mu^{\star},\nu_0}\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right) - \frac{1}{\gamma^2}V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}} + \frac{2}{\gamma^2\left(1-\gamma\right)}\big(I-\gamma P^{\mu^{\star},\nu_0}\big)V^{-}_{\mathsf{pe}} + \frac{4}{\gamma^2\left(1-\gamma\right)}\beta^{\mu^{\star},\nu_0} \\
&\overset{\text{(ii)}}{\leq} P^{\mu^{\star},\nu_0}\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right) - \frac{1}{\gamma}V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}} + \frac{2}{\gamma^2\left(1-\gamma\right)}\big(I-\gamma P^{\mu^{\star},\nu_0}\big)V^{-}_{\mathsf{pe}} + \frac{4}{\gamma^2\left(1-\gamma\right)}\beta^{\mu^{\star},\nu_0} \\
&= \big(I-\gamma P^{\mu^{\star},\nu_0}\big)\left[\frac{2}{\gamma^2\left(1-\gamma\right)}V^{-}_{\mathsf{pe}} - \frac{1}{\gamma}\left(V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}}\right)\right] + \frac{4}{\gamma^2\left(1-\gamma\right)}\beta^{\mu^{\star},\nu_0},
\end{aligned}
$$

where (ii) holds since $\gamma < 1$. To see why (i) holds, we make the following observation:

$$
\begin{aligned}
V^{-}_{\mathsf{pe}}\circ V^{-}_{\mathsf{pe}} - \gamma^2\big(P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\big)\circ\big(P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\big) &= \big(V^{-}_{\mathsf{pe}} - \gamma P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\big)\circ\big(V^{-}_{\mathsf{pe}} + \gamma P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\big) \\
&\leq \big(V^{-}_{\mathsf{pe}} - \gamma P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}} + 2\beta^{\mu^{\star},\nu_0}\big)\circ\big(V^{-}_{\mathsf{pe}} + \gamma P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}}\big) \\
&\leq \frac{2}{1-\gamma}\big(I-\gamma P^{\mu^{\star},\nu_0}\big)V^{-}_{\mathsf{pe}} + \frac{4}{1-\gamma}\beta^{\mu^{\star},\nu_0},
\end{aligned}
$$

where the penultimate line arises from $\beta^{\mu^{\star},\nu_0} \geq 0$ and $V^{-}_{\mathsf{pe}} + \gamma P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}} \geq 0$, and the last line holds since $V^{-}_{\mathsf{pe}} - \gamma P^{\mu^{\star},\nu_0}V^{-}_{\mathsf{pe}} + 2\beta^{\mu^{\star},\nu_0} \geq 0$ (due to (B.16)) and $\widehat{V}_{\mathsf{pe}} + \gamma P^{\mu^{\star},\nu_0}\widehat{V}_{\mathsf{pe}} \leq \frac{2}{1-\gamma}1$.

Making use of the following identity (see (B.20))

$$\left(d^{\mu^\star,\nu_0}\right)^\top = (1-\gamma)\,\rho^\top\left(I-\gamma P^{\mu^\star,\nu_0}\right)^{-1},$$

we can deduce that

$$\begin{aligned}
\sum_{s\in\mathcal{S}} d^{\mu^\star,\nu_0}(s;\rho)\,\mathsf{Var}_{P^{\mu^\star,\nu_0}(\cdot\,|\,s)}\left(V_{\mathsf{pe}}^-\right) &= \left(d^{\mu^\star,\nu_0}\right)^\top v \\
&\leq (1-\gamma)\,\rho^\top\left[\frac{2}{\gamma^2(1-\gamma)}V_{\mathsf{pe}}^- - \frac{1}{\gamma}\left(V_{\mathsf{pe}}^-\circ V_{\mathsf{pe}}^-\right)\right] + \frac{4}{\gamma^2(1-\gamma)}\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0} \\
&\leq \frac{2}{\gamma^2}\rho^\top V_{\mathsf{pe}}^- + \frac{4}{\gamma^2(1-\gamma)}\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0} \\
&\leq \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)}\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0}, \qquad \text{(C.23)}
\end{aligned}$$

where the last line holds since $\rho^\top V_{\mathsf{pe}}^- \leq \|V_{\mathsf{pe}}^-\|_\infty \leq \frac{1}{1-\gamma}$. Taking (C.22) and (C.23) collectively implies that

$$\begin{aligned}
\alpha_3 &\leq 2\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N}\log\frac{N}{(1-\gamma)\delta}}\sqrt{\frac{2}{\gamma^2(1-\gamma)}+\frac{4}{\gamma^2(1-\gamma)}\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0}} \\
&\leq 8\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)}\log\frac{N}{(1-\gamma)\delta}}\left(1+\sqrt{\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0}}\right), \qquad \text{(C.24)}
\end{aligned}$$

where the last relation holds from the assumption $\gamma \geq 1/2$ and the elementary inequality $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$.

Putting all this together. Taking together the above bounds (C.20), (C.21) and (C.24) (on α_1, α_2 and α_3, respectively) leads to

$$\begin{aligned}
\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0} &\leq c_3\alpha_1 + c_3\alpha_2 + \frac{4}{N} \leq (c_3+c_5)\,\alpha_1 + c_5\alpha_3 + \frac{4}{N} \\
&\leq (2c_3+2c_5+4)\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta} \\
&\quad + 8c_5\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)}\log\frac{N}{(1-\gamma)\delta}}\left(1+\sqrt{\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0}}\right),
\end{aligned}$$

where we have used the fact that $\frac{1}{N}\leq\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta}$. In turn, this implies that there exists some sufficiently large constant $\widetilde{C}>0$ such that

$$\begin{aligned}
\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0} &\leq \widetilde{C}\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta} + \widetilde{C}\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)}\log\frac{N}{(1-\gamma)\delta}}\left(1+\sqrt{\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0}}\right) \\
&\leq \widetilde{C}\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta} + \widetilde{C}\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)}\log\frac{N}{(1-\gamma)\delta}} \\
&\quad + \frac{\widetilde{C}^2}{2}\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)}\log\frac{N}{(1-\gamma)\delta} + \frac{1}{2}\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0},
\end{aligned}$$

where the last relation follows from the AM-GM inequality. Rearranging terms, we are left with

$$\left(d^{\mu^\star,\nu_0}\right)^\top\beta^{\mu^\star,\nu_0} \leq \left(2\widetilde{C}+\widetilde{C}^2\right)\frac{C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)N}\log\frac{N}{(1-\gamma)\delta} + 2\widetilde{C}\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)}\log\frac{N}{(1-\gamma)\delta}},$$

thus concluding the proof.

C.6.1 Proof of Claim 1

For any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, if $N d_{\mathsf{b}}(s,a,b) \leq 2\log(N/\delta)$, then the claim (C.18) holds trivially. It thus suffices to show that with probability exceeding $1-\delta$, (C.18) holds for all (s,a,b) falling within the following set

$$\mathcal{N}_{\mathsf{large}} := \left\{(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} : N d_{\mathsf{b}}(s,a,b) \geq 2\log\frac{N}{\delta}\right\}.$$

It is straightforward to check that the cardinality of $\mathcal{N}_{\mathsf{large}}$ obeys

$$|\mathcal{N}_{\mathsf{large}}| \cdot \frac{2}{N}\log\frac{N}{\delta} = \sum_{(s,a,b)\in\mathcal{N}_{\mathsf{large}}} \frac{2}{N}\log\frac{N}{\delta} \leq \sum_{(s,a,b)\in\mathcal{N}_{\mathsf{large}}} d_{\mathsf{b}}(s,a,b) \leq 1,$$

and as a result,

$$|\mathcal{N}_{\mathsf{large}}| \leq \frac{N}{2\log(N/\delta)} \leq \frac{N}{2}.$$

In view of the Chernoff bound (Vershynin, 2018, Exercise 2.3.2), for any $(s,a,b) \in \mathcal{N}_{\mathsf{large}}$, we have

$$\begin{aligned}\mathbb{P}\left(N(s,a,b) \geq \frac{1}{2}N d_{\mathsf{b}}(s,a,b)\right) &\leq e^{-N d_{\mathsf{b}}(s,a,b)}\left(\frac{e}{2}\right)^{\frac{1}{2}N d_{\mathsf{b}}(s,a,b)} = \left(\frac{1}{2e}\right)^{\frac{1}{2}N d_{\mathsf{b}}(s,a,b)} \\ &\leq \left(\frac{1}{2e}\right)^{\log(N/\delta)} \leq \frac{\delta}{N}.\end{aligned}$$

Invoke the union bound to reach

$$\mathbb{P}\left(N(s,a,b) \geq \frac{1}{2}N d_{\mathsf{b}}(s,a,b), \forall (s,a,b) \in \mathcal{N}_{\mathsf{large}}\right) \leq \frac{\delta}{N}|\mathcal{N}_{\mathsf{large}}| \leq \delta.$$

Putting the above two cases together completes the proof.

C.6.2 Proof of Claim 2

In view of Lemma 2, we know that with probability exceeding $1-\delta$,

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}(\widetilde{V}) \leq 2\mathsf{Var}_{P_{s,a,b}}(\widetilde{V}) + O\left(\frac{\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a,b)}\right) \tag{C.25}$$

holds simultaneously for all $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ satisfying $N(s,a,b) \geq 1$. As a result, with probability exceeding $1-\delta$,

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}(\widetilde{V}) \leq 2\mathsf{Var}_{P_{s,a,b}}(\widetilde{V}) + O\left(\frac{\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 \max\left\{N(s,a,b), \log\frac{N}{\delta}\right\}}\right) \tag{C.26}$$

holds simultaneously for all $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. To see why this is valid, let us divide into two cases:

1. When $N(s,a,b) \geq \log\frac{N}{\delta}$, the relation (C.26) is clearly guaranteed by (C.25);
2. When $N(s,a,b) \leq \log\frac{N}{\delta}$, then it is seen that

$$\begin{aligned}\mathsf{Var}_{\widehat{P}_{s,a,b}}(\widetilde{V}) &\leq \frac{1}{(1-\gamma)^2} = O\left(\frac{\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 \max\left\{N(s,a,b), \log\frac{N}{\delta}\right\}}\right) \\ &\leq 2\mathsf{Var}_{P_{s,a,b}}(\widetilde{V}) + O\left(\frac{\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 \max\left\{N(s,a,b), \log\frac{N}{\delta}\right\}}\right),\end{aligned}$$

   which again yields (C.26).

Combine (C.26) with Claim 1 to establish the desired bound:

$$\mathsf{Var}_{\widehat{P}_{s,a,b}}(V_{\mathsf{pe}}^{-}) \leq 2\mathsf{Var}_{P_{s,a,b}}(V_{\mathsf{pe}}^{-}) + O\left(\frac{\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N d_{\mathsf{b}}(s,a,b)}\right).$$

D Proof of Theorem 2

This section presents the proof of the minimax lower bound stated in Theorem 2. Without loss of generality, we assume throughout the proof that $A \geq B$.

D.1 Constructing a family of hard Markov game instances

The first step lies in constructing a family of Markov games $\{\mathcal{MG}_\theta\}$ each parameterized by a $\theta \in \{p,q\}^A$, where

$$p \coloneqq \gamma + 14\frac{(1-\gamma)^2\varepsilon}{\gamma} \qquad \text{and} \qquad q \coloneqq \gamma - 14\frac{(1-\gamma)^2\varepsilon}{\gamma}. \tag{D.1}$$

Clearly, under the conditions $\gamma \geq 2/3$ and $0 < \varepsilon \leq \frac{1}{42(1-\gamma)}$, it holds that

$$\frac{1}{2} \leq \gamma - \frac{1-\gamma}{2} \leq q < p \leq \gamma + \frac{1-\gamma}{2} \leq 1. \tag{D.2}$$

For each $\theta = \{\theta_a\}_{a\in\mathcal{A}} \in \{p,q\}^A$, we define the corresponding Markov game $\mathcal{MG}_\theta$, and the associated initial state distribution $\rho \in \Delta(\mathcal{S})$ and the data distribution $d_{\mathsf{b}} \in \Delta(\mathcal{S}\times\mathcal{A}\times\mathcal{B})$ as follows:

- Let the state space be $\mathcal{S} = \{0,1,2,\dots,S-1\}$, and the action spaces for the two players be $\mathcal{A} = \{0,1,\dots,A-1\}$ and $\mathcal{B} = \{0,1,\dots,B-1\}$, respectively.
- Define the transition kernel $P_\theta : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \Delta(\mathcal{S})$ such that

  $$P_\theta\left(s' \mid s,a,b\right) = \begin{cases} \theta_a \mathbb{1}\left\{s'=0\right\} + (1-\theta_a)\mathbb{1}\left\{s'=1\right\}, & \text{if } s=0 \text{ and } b=0 \\ \mathbb{1}\left\{s'=s\right\}, & \text{if } s\geq 1 \text{ or } b \geq 1 \end{cases} \tag{D.3}$$

  for any $s,s' \in \mathcal{S}$, $a\in\mathcal{A}$ and $b \in \mathcal{B}$.
- Set the reward function to be

  $$r\left(s,a,b\right) = \mathbb{1}\left\{s=0\right\} \qquad \text{for all } (s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}. \tag{D.4}$$

- The initial state distribution $\rho\in\Delta(\mathcal{S})$ is taken to be

  $$\rho\left(s\right) = \mathbb{1}\left\{s=0\right\}, \qquad \text{for all}\, s\in\mathcal{S}. \tag{D.5}$$

- The distribution $d_{\mathsf{b}} \in \Delta(\mathcal{S}\times\mathcal{A}\times\mathcal{B})$ generating the batch dataset is defined such that

  $$d_{\mathsf{b}}\left(s,a,b\right) = \begin{cases} \frac{1}{C^\star_{\mathsf{clipped}}S(A+B)}, & \text{if } s=0 \\ \frac{1}{AB}\left(1-\frac{AB}{C^\star_{\mathsf{clipped}}S(A+B)}\right), & \text{if } s=1 \\ 0, & \text{if } s\geq 2 \end{cases} \tag{D.6}$$

  for all $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, where $C^\star_{\mathsf{clipped}}$ is allowed to be any given quantity satisfying $C^\star_{\mathsf{clipped}} \geq \frac{2AB}{S(A+B)}$. In other words, for any given $s\in\mathcal{S}$, each action pair $(a,b)\in\mathcal{A}\times\mathcal{B}$ is sampled with the same rate.

Owing to the simple structure of the above $\mathcal{MG}_\theta$, we are able to compute its value functions and the Nash equilibrium. Before proceeding, we find it convenient to define two action subsets

$$\mathcal{A}_{p,\theta} \coloneqq \{a\in\mathcal{A} : \theta_a = p\} \qquad \text{and} \qquad \mathcal{A}_{q,\theta} \coloneqq \{a\in\mathcal{A} : \theta_a = q\}. \tag{D.7}$$

Lemma 6. *Consider the Markov game $\mathcal{MG}_\theta$. For any policy pair (μ,ν), the value function in $\mathcal{MG}_\theta$ obeys*

$$V^{\mu,\nu}\left(0\right) = \frac{1}{1-\gamma+\gamma\mu_{p,\theta}\nu_0\left(1-p\right)+\gamma\left(1-\mu_{p,\theta}\right)\nu_0\left(1-q\right)},$$
$$V^{\mu,\nu}\left(s\right) = 0 \qquad \textit{for all } s\geq 1,$$

where

$$\mu_{p,\theta} := \sum_{a\in\mathcal{A}_{p,\theta}} \mu(a\,|\,0) \qquad \textit{and} \qquad \nu_0 := \nu(0\,|\,0). \tag{D.8}$$

In addition, if we define two policies $\mu_\theta^\star$ and $\nu_\theta^\star$ as follows

$$\mu_\theta^\star\left(a\,|\,s\right) = \frac{1}{|\mathcal{A}_{p,\theta}|}\mathbb{1}\left\{a\in\mathcal{A}_{p,\theta}\right\}, \qquad \forall\, s\in\mathcal{S}, a\in\mathcal{A}, \tag{D.9a}$$

$$\nu_\theta^\star\left(b\,|\,s\right) = \mathbb{1}\left\{b=0\right\}, \qquad \forall\, s\in\mathcal{S}, b\in\mathcal{B}, \tag{D.9b}$$

then the policy pair $(\mu_\theta^\star, \nu_\theta^\star)$ is a Nash equilibrium of $\mathcal{MG}_\theta$. Furthermore, we have

$$\left\{\mathcal{MG}_\theta, \rho, d_{\mathsf{b}}\right\} \in \mathsf{MG}\big(C^\star_{\mathsf{clipped}}\big).$$

Proof. See Appendix E.1. ☐

We shall also make note of two immediate consequences of Lemma 6. First, Lemma 6 allows us to compute the corresponding value function $V^\star$ of the unique Nash equilibrium in $\mathcal{MG}_\theta$ as follows:

$$V^\star\left(0\right) = \frac{1}{1-\gamma p}, \tag{D.10a}$$

given that $\mu_{p,\theta}=1$ and $\nu_0=1$ hold under the NE $(\mu_\theta^\star, \nu_\theta^\star)$. In addition, Lemma 6 tells us that for any policy μ of the max-player, the optimal value function for the min-player obeys

$$V^{\mu,\star}(0) = \frac{1}{1-\gamma p+\gamma\left(1-\mu_{p,\theta}\right)\left(p-q\right)}, \tag{D.10b}$$

given that the best response of the min-player is to set $\nu_0=1$.

D.2 Identifying sufficient statistics for the parameter θ

Denote by f_θ the probability mass function of a single sample transition (s_i, a_i, b_i, s_i') such that

$$(s_i, a_i, b_i) \sim d_{\mathsf{b}}, \qquad s_i' \sim P_\theta\left(\cdot\,|\,s_i, a_i, b_i\right).$$

In view of the definition of d_{b} in (D.6), we can write

$$f_\theta\left(s_i,a_i,b_i,s_i'\right) = \begin{cases} \frac{1}{C^\star_{\mathsf{clipped}}S(A+B)}\theta_{a_i} & \text{if } s_i=0, b_i=0, s_i'=0,\\ \frac{1}{C^\star_{\mathsf{clipped}}S(A+B)}\left(1-\theta_{a_i}\right) & \text{if } s_i=0, b_i=0, s_i'=1,\\ \frac{1}{C^\star_{\mathsf{clipped}}S(A+B)} & \text{if } s_i=0, b_i\geq 1, s_i'=0,\\ \frac{1}{AB}\left(1-\frac{AB}{C^\star_{\mathsf{clipped}}S(A+B)}\right) & \text{if } s_i=s_i'=1,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore, the probability mass function of the offline dataset $\{(s_i,a_i,b_i,s_i')\}_{i=1}^N$ containing N independent data is given by

$$\begin{aligned}\prod_{i=1}^N f_\theta\left(s_i,a_i,b_i,s_i'\right) &= \left[\frac{1}{AB}\left(1-\frac{AB}{C^\star_{\mathsf{clipped}}S\left(A+B\right)}\right)\right]^L\left[\frac{1}{C^\star_{\mathsf{clipped}}S\left(A+B\right)}\right]^J\\ &\quad\cdot\prod_{a\in\mathcal{A}}\left[\frac{1}{C^\star_{\mathsf{clipped}}S\left(A+B\right)}\theta_a\right]^{X_a}\left[\frac{1}{C^\star_{\mathsf{clipped}}S\left(A+B\right)}\left(1-\theta_a\right)\right]^{M_a-X_a}\\ &= \left[\frac{1}{AB}\left(1-\frac{AB}{C^\star_{\mathsf{clipped}}S\left(A+B\right)}\right)\right]^L\left[\frac{1}{C^\star_{\mathsf{clipped}}S\left(A+B\right)}\right]^{N+J-L}\prod_{a\in\mathcal{A}}\theta_a^{X_a}\left(1-\theta_a\right)^{M_a-X_a},\end{aligned} \tag{D.11}$$

where we define

$$J \coloneqq \sum_{i=1}^{N} \mathbb{1}\{s_i = s_i' = 0, b_i \geq 1\},$$
$$L \coloneqq \sum_{i=1}^{N} \mathbb{1}\{s_i = s_i' = 1\},$$
$$X_a \coloneqq \sum_{i=1}^{N} \mathbb{1}\left\{s_i = 0, a_i = a, b_i = 0, s_i' = 0\right\}, \qquad \forall a \in \mathcal{A},$$
$$M_a \coloneqq \sum_{i=1}^{N} \mathbb{1}\left\{s_i = 0, a_i = a, b_i = 0\right\}, \qquad \forall a \in \mathcal{A}.$$

By the factorization theorem (see, e.g., Shao (2003, Theorem 2.2)), the expression (D.11) implies that

$$(X, M) \qquad \text{with } X = [X_a]_{a\in\mathcal{A}} \text{ and } M = [M_a]_{a\in\mathcal{A}} \tag{D.12}$$

forms a sufficient statistic for the underlying parameter θ.

D.3 Establishing the minimax lower bound

In view of (D.10) and our choice of ρ (cf. (D.5)), we can obtain — for any policy μ of the max-player — that

$$\begin{aligned}
V^{\star}(\rho) - V^{\mu,\star}(\rho) = V^{\star}(0) - V^{\mu,\star}(0) &= \frac{1}{1-\gamma p} - \frac{1}{1-\gamma p + \gamma\left(1-\mu_{p,\theta}\right)(p-q)} \\
&= \frac{\gamma\left(1-\mu_{p,\theta}\right)(p-q)}{(1-\gamma p)\left[1-\gamma p + \gamma\left(1-\mu_{p,\theta}\right)(p-q)\right]} \\
&= \left(1-\mu_{p,\theta}\right)\frac{28\left(1-\gamma\right)^2\varepsilon}{\left[1-\gamma\left(\gamma+14\frac{(1-\gamma)^2\varepsilon}{\gamma}\right)\right]\left[1-\gamma\left(\gamma+14\frac{(1-\gamma)^2\varepsilon}{\gamma}\right)+\left(1-\mu_{p,\theta}\right)28\left(1-\gamma\right)^2\varepsilon\right]} \\
&= \left(1-\mu_{p,\theta}\right)\frac{28\varepsilon}{\left[1+\gamma-14\left(1-\gamma\right)\varepsilon\right]\left[1+\gamma+\left(1-2\mu_{p,\theta}\right)14\left(1-\gamma\right)\varepsilon\right]} \\
&\geq \left(1-\mu_{p,\theta}\right)6\varepsilon = 6\varepsilon\sum_{a\in\mathcal{A}_{q,\theta}}\mu\left(a\,|\,0\right) = 6\varepsilon\sum_{a\in\mathcal{A}}\mu\left(a\,|\,0\right)\mathbb{1}\left\{\theta_a = q\right\},
\end{aligned} \tag{D.13}$$

where the penultimate relation holds provided that $\varepsilon \leq \frac{1}{42(1-\gamma)}$. As a consequence, we have

$$\begin{aligned}
\inf_{(\widehat{\mu},\widehat{\nu})}\sup_{\{\mathcal{MG},\rho,d_{\mathsf{b}}\}\in\mathsf{MG}(C^{\star}_{\mathsf{clipped}})}\mathbb{E}\left[V^{\star,\widehat{\nu}}(\rho) - V^{\widehat{\mu},\star}(\rho)\right] &\geq \inf_{\widehat{\mu}}\sup_{\{\mathcal{MG},\rho,d_{\mathsf{b}}\}\in\mathsf{MG}(C^{\star}_{\mathsf{clipped}})}\mathbb{E}\left[V^{\star}(\rho) - V^{\widehat{\mu},\star}(\rho)\right] \\
&\geq \inf_{\widehat{\mu}}\sup_{\mathcal{MG}_{\theta}:\theta\in\{p,q\}^A}\mathbb{E}\left[V^{\star}(\rho) - V^{\widehat{\mu},\star}(\rho)\right] \\
&\geq 6\varepsilon\inf_{\widehat{\mu}}\sup_{\mathcal{MG}_{\theta}:\theta\in\{p,q\}^A}\mathbb{E}\left[\sum_{a\in\mathcal{A}}\widehat{\mu}\left(a\,|\,0\right)\mathbb{1}\left\{\theta_a = q\right\}\right],
\end{aligned} \tag{D.14}$$

where the first relation holds since $V^{\star,\widehat{\nu}}(s) \geq V^{\mu^{\star},\widehat{\nu}}(s) \geq V^{\mu^{\star},\nu^{\star}}(s) = V^{\star}(s)$, and the last line follows from (D.13). Here, the expectation is taken over the randomness of the data distribution induced by d_{b} and that of the transition kernel of the corresponding Markov game.

To further lower bound (D.14), we define $\mathcal{U}$ to be the uniform distribution over $\{p,q\}^A$ (namely, a random vector $\theta \sim \mathcal{U}$ obeys $\theta_a \overset{\text{i.i.d.}}{=} \begin{cases} p & \text{w.p. } 0.5 \\ q & \text{w.p. } 0.5 \end{cases}$ for all $a \in \mathcal{A}$). It is readily seen that

$$\inf_{\widehat{\mu}}\sup_{\mathcal{MG}_{\theta}:\theta\in\{p,q\}^A}\mathbb{E}\left[\sum_{a\in\mathcal{A}}\widehat{\mu}\left(a\,|\,0\right)\mathbb{1}\left\{\theta_a = q\right\}\right] \overset{\text{(i)}}{\geq} \inf_{\widehat{\mu}}\mathbb{E}_{\theta\sim\mathcal{U}}\left[\mathbb{E}_{\mathcal{MG}_{\theta}}\left[\sum_{a\in\mathcal{A}}\widehat{\mu}\left(a\,|\,0\right)\mathbb{1}\left\{\theta_a = q\right\}\right]\right]$$

$$= \inf_{\widehat{\mu}} \mathop{\mathbb{E}}_{\mathcal{MG}_\theta, \theta\sim\mathcal{U}} \left[\mathbb{E}\left[\sum_{a\in\mathcal{A}} \widehat{\mu}\left(a \,|\, 0\right) \mathbb{1}\left\{\theta_a = q\right\} \mid X, M \right]\right]$$

$$\overset{\text{(ii)}}{=} \inf_{\widehat{\mu}} \mathop{\mathbb{E}}_{\mathcal{MG}_\theta, \theta\sim\mathcal{U}} \left[\sum_{a\in\mathcal{A}} \widehat{\mu}\left(a \,|\, 0\right) \mathbb{E}\Big[\mathbb{1}\left\{\theta_a = q\right\} \mid X, M \Big]\right]$$

$$\overset{\text{(iii)}}{=} \inf_{\widehat{\mu}} \mathop{\mathbb{E}}_{\mathcal{MG}_\theta, \theta\sim\mathcal{U}} \left[\sum_{a\in\mathcal{A}} \widehat{\mu}\left(a \,|\, 0\right) \mathbb{P}\left(\theta_a = q \,|\, X_a, M_a\right)\right], \tag{D.15}$$

where the expectations appear after (i) are taken w.r.t. the randomness of data distribution induced by d_{b} and the transition kernel of $\mathcal{MG}_\theta$, with θ following a prior distribution $\mathcal{U}$. To see why (ii) holds, note that it suffices to consider $\widehat{\mu}$ that is a measurable function of the sufficient statistic (X, M) (see Johnstone (2017, Proposition 3.13)); with regards (iii), we note that it holds since $\{X_{a'}, M_{a'}\}_{a'\neq a}$ is independent of θ_a conditional on (X_a, M_a). To further lower bound (D.15), we look at the following two cases separately.

Case 1: when the sample size N is not too small. Consider the case where

$$N \geq \widetilde{C} \frac{S\left(A+B\right) C^{\star}_{\mathsf{clipped}} \log\left(A+B\right)}{1-\gamma} = \frac{1}{d_{\mathsf{b}}\left(0,a,0\right)} \cdot \frac{\widetilde{C}\log\left(A+B\right)}{1-\gamma} \tag{D.16}$$

holds for some sufficiently large constant $\widetilde{C} > 0$. In order to be compatible with the assumption

$$N < \frac{c_2 S\left(A+B\right) C^{\star}_{\mathsf{clipped}}}{\left(1-\gamma\right)^3 \varepsilon^2 \log\left(A+B\right)} = \frac{1}{d_{\mathsf{b}}\left(0,a,0\right)} \cdot \frac{c_2}{\left(1-\gamma\right)^3 \varepsilon^2 \log\left(A+B\right)}, \tag{D.17}$$

it suffices to focus on the regime where

$$\varepsilon \leq \varepsilon_0 \coloneqq \sqrt{\frac{c_2}{2\widetilde{C}}} \frac{1}{\left(1-\gamma\right)\log\left(A+B\right)}. \tag{D.18}$$

In order to understand (D.15), we need to take a look at M_a and X_a. To begin with, it is straightforward to check that $M_a \sim \mathsf{Binomial}\big(N, d_{\mathsf{b}}(0,a,0)\big)$. When N is sandwiched between (D.16) and (D.17), we can establish the following high-probability bound on M_a.

Lemma 7. *Suppose that $\widetilde{C}$ is sufficiently large, and $\varepsilon \leq \frac{1}{(1-\gamma)\log A}$. Then we have*

$$\mathbb{P}\left(\frac{c_3 \log A}{1-\gamma} \leq M_a \leq \frac{c_4}{\left(1-\gamma\right)^3 \varepsilon^2 \log A}\right) \geq 1 - \frac{2}{A^4}, \tag{D.19}$$

where $c_3 = \widetilde{C}/2$ and $c_4 = c_2 + \sqrt{8c_2/c_{\mathsf{ch}}}$, with $c_{\mathsf{ch}} > 0$ some universal constant independent of c_2.

Proof. See Appendix E.2. □

In words, M_a is guaranteed to reside within an interval $\big[\frac{c_3\log A}{1-\gamma}, \frac{c_4}{(1-\gamma)^3\varepsilon^2\log A}\big]$. In addition, in order to bound $\mathbb{P}\left(\theta_a = q \,|\, X_a, M_a\right)$ in (D.15), we seek to first investigate the properties of $X_a \mid \{\theta_a, M_a = M\}$ for some $M \in \big[\frac{c_3\log A}{1-\gamma}, \frac{c_4}{(1-\gamma)^3\varepsilon^2\log A}\big]$. Clearly, conditional on θ_a and $M_a = M$, the random variable X_a is distributed as $\mathsf{Binomial}(M, \theta_a)$. We make note of the following useful result regarding binomial random variables.

Lemma 8. *Suppose that $\widetilde{C}$ is sufficiently large and c_2 is sufficiently small. Consider any integer M satisfying*

$$\frac{c_3 \log A}{1-\gamma} \leq M \leq \frac{c_4}{\left(1-\gamma\right)^3 \varepsilon^2 \log A}, \tag{D.20}$$

and generate $B_p \sim \mathsf{Binomial}(M, p)$ and $B_q \sim \mathsf{Binomial}(M, q)$. Then there exists a set $E_M \subseteq [M]$ such that

$$\mathbb{P}\left(B_p \in E_M\right) \geq 1 - \frac{2}{A^4}, \qquad \mathbb{P}\left(B_q \in E_M\right) \geq 1 - \frac{2}{A^4}, \tag{D.21}$$

$$\text{and} \qquad \frac{\mathbb{P}\left(B_q = n\right)}{\mathbb{P}\left(B_p = n\right)} \geq \frac{1}{2}, \qquad \forall n \in E_M. \tag{D.22}$$

Proof. See Appendix E.3. $\square$

In words, when M falls within the range (D.20), there exists a high-probability set E_M such that it is not that easy to differentiate $\mathsf{Binomial}(M,p)$ and $\mathsf{Binomial}(M,q)$. This result motivates us to pay particular attention the following set $E \subseteq \mathbb{N}^2$:

$$E := \left\{ (x,m) \in \mathbb{N}^2 : x \in E_m, \frac{c_3 \log A}{1-\gamma} \leq m \leq \frac{c_4}{(1-\gamma)^3 \varepsilon^2 \log A} \right\}.$$

It is seen that

$$\begin{aligned}
\mathbb{P}\Big((X_a, M_a) \in E \,|\, \theta_a = p \Big) &= \mathbb{E}\Big[\mathbb{P}\big((X_a, M_a) \in E \,|\, M_a, \theta_a = p \big) \,|\, \theta_a = p \Big] = \mathbb{E}\Big[\mathbb{P}\left(X_a \in E_{M_a} \,|\, M_a, \theta_a = p \right) \,|\, \theta_a = p \Big] \\
&\geq \mathbb{E}\left[\mathbb{P}\left(X_a \in E_{M_a} \,|\, M_a, \theta_a = p \right) \mathbb{1}\left\{ \frac{c_3 \log A}{1-\gamma} \leq M_a \leq \frac{c_4}{(1-\gamma)^3 \varepsilon^2 \log A} \right\} \,|\, \theta_a = p \right] \\
&\overset{\text{(i)}}{\geq} \mathbb{E}\left[\left(1 - \frac{2}{A^4} \right) \mathbb{1}\left\{ \frac{c_3 \log A}{1-\gamma} \leq M_a \leq \frac{c_4}{(1-\gamma)^3 \varepsilon^2 \log A} \right\} \,|\, \theta_a = p \right] \\
&\overset{\text{(ii)}}{=} \left(1 - \frac{2}{A^4} \right) \mathbb{P}\left(\frac{c_3 \log A}{1-\gamma} \leq M_a \leq \frac{c_4}{(1-\gamma)^3 \varepsilon^2 \log A} \right) \\
&\overset{\text{(iii)}}{\geq} \left(1 - \frac{2}{A^4} \right)^2 \geq 1 - \frac{4}{A^4}. \qquad \text{(D.23)}
\end{aligned}$$

Here, (i) utilizes Lemma 8, (ii) is valid since θ_a is independent of M_a, whereas (iii) follows from (E.9) and (E.10). Similarly, one can show that

$$\mathbb{P}\Big((X_a, M_a) \in E \,|\, \theta_a = q \Big) \geq 1 - \frac{4}{A^4}. \qquad \text{(D.24)}$$

Taking (D.23) and (D.24) collectively yields

$$\begin{aligned}
\mathbb{P}\Big((X_a, M_a) \in E \Big) &= \mathbb{P}\Big((X_a, M_a) \in E \,|\, \theta_a = p \Big) \mathbb{P}\left(\theta_a = p \right) + \mathbb{P}\Big((X_a, M_a) \in E \,|\, \theta_a = q \Big) \mathbb{P}\left(\theta_a = q \right) \\
&\geq \frac{1}{2}\left(1 - \frac{4}{A^4} \right) + \frac{1}{2}\left(1 - \frac{4}{A^4} \right) = 1 - \frac{4}{A^4}. \qquad \text{(D.25)}
\end{aligned}$$

A little calculation shows that: when $(x_a, m_a) \in E$, we have

$$\begin{aligned}
&\mathbb{P}\left(\theta_a = q \,|\, X_a = x_a, M_a = m_a \right) = \frac{\mathbb{P}\left(\theta_a = q, X_a = x_a \,|\, M_a = m_a \right)}{\mathbb{P}\left(X_a = x_a \,|\, M_a = m_a \right)} \\
&= \frac{\mathbb{P}\left(X_a = x_a \,|\, \theta_a = q, M_a = m_a \right) \mathbb{P}\left(\theta_a = q \,|\, M_a = m_a \right)}{\mathbb{P}\left(X_a = x_a \,|\, \theta_a = p, M_a = m_a \right) \mathbb{P}\left(\theta_a = p \,|\, M_a = m_a \right) + \mathbb{P}\left(X_a = x_a \,|\, \theta_a = q, M_a = m_a \right) \mathbb{P}\left(\theta_a = q \,|\, M_a = m_a \right)} \\
&\overset{\text{(i)}}{=} \frac{\mathbb{P}\left(X_a = x_a \,|\, \theta_a = q, M_a = m_a \right) \mathbb{P}\left(\theta_a = q \right)}{\mathbb{P}\left(X_a = x_a \,|\, \theta_a = p, M_a = m_a \right) \mathbb{P}\left(\theta_a = p \right) + \mathbb{P}\left(X_a = x_a \,|\, \theta_a = q, M_a = m_a \right) \mathbb{P}\left(\theta_a = q \right)} \\
&\overset{\text{(ii)}}{=} \frac{\mathbb{P}\left(X_a = x_a \,|\, \theta_a = q, M_a = m_a \right)}{\mathbb{P}\left(X_a = x_a \,|\, \theta_a = p, M_a = m_a \right) + \mathbb{P}\left(X_a = x_a \,|\, \theta_a = q, M_a = m_a \right)} \\
&\overset{\text{(iii)}}{=} \frac{\mathbb{P}\left(B_q = x_a \right)}{\mathbb{P}\left(B_p = x_a \right) + \mathbb{P}\left(B_q = x_a \right)} \overset{\text{(iv)}}{\geq} \frac{1/2}{1 + 1/2} = \frac{1}{3}, \qquad \text{(D.26)}
\end{aligned}$$

where we let B_p and B_q be two random variables distributed as $B_p \sim \mathsf{Binomial}(m_a, p)$ and $B_q \sim \mathsf{Binomial}(m_a, q)$. Here, (i) relies on the fact that θ_a is independent from M_a; (ii) comes from the fact that $\mathbb{P}\left(\theta_a = p \right) = \mathbb{P}\left(\theta_a = q \right) = 1/2$; (iii) follows since conditional on $M_a = m_a$ and $\theta_a = p$ (resp. $\theta_a = q$), the distribution of X_a is $\mathsf{Binomial}(m_a, p)$ (resp. $\mathsf{Binomial}(m_a, q)$); and (iv) follows from Lemma 8 given that $(x_a, m_a) \in E$. We can thus conclude that

$$\inf_{(\widehat{\mu}, \widehat{\nu})} \sup_{\{\mathcal{MG}, \rho, d_{\mathsf{b}}\} \in \mathsf{MG}(C^\star)} \mathbb{E}\left[V^{\star, \widehat{\nu}}(\rho) - V^{\widehat{\mu}, \star}(\rho) \right]$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\geq} 6\varepsilon \inf_{\widehat{\mu}} \mathbb{E}\left[\sum_{a\in\mathcal{A}} \widehat{\mu}\left(a \,|\, 0\right) \mathbb{P}\left(\theta_a = q \,|\, X_a, M_a\right)\right] \\
&\geq 6\varepsilon \inf_{\widehat{\mu}} \mathbb{E}\left[\sum_{a\in\mathcal{A}} \widehat{\mu}\left(a \,|\, 0\right) \mathbb{P}\left(\theta_a = q \,|\, X_a, M_a\right) \mathbb{1}\left\{(X_a, M_a) \in E\right\}\right] \\
&\overset{\text{(ii)}}{\geq} 2\varepsilon \inf_{\widehat{\mu}} \mathbb{E}\left[\sum_{a\in\mathcal{A}} \widehat{\mu}\left(a \,|\, 0\right) \mathbb{1}\left\{(X_a, M_a) \in E\right\}\right] \\
&= 2\varepsilon \inf_{\widehat{\mu}} \mathbb{E}\left[1 - \sum_{a\in\mathcal{A}} \widehat{\mu}\left(a \,|\, 0\right) \mathbb{1}\left\{(X_a, M_a) \notin E\right\}\right] \\
&\geq 2\varepsilon\, \mathbb{E}\left[1 - \sum_{a\in\mathcal{A}} \mathbb{1}\left\{(X_a, M_a) \notin E\right\}\right] = 2\varepsilon\left[1 - \sum_{a\in\mathcal{A}} \mathbb{P}\Big(\left(X_a, M_a\right) \notin E\Big)\right] \\
&\overset{\text{(iii)}}{\geq} 2\varepsilon\left(1 - \frac{4}{A^3}\right) \overset{\text{(iv)}}{\geq} \varepsilon.
\end{aligned}
$$

Here, (i) follows from (D.14) and (D.15); (ii) makes use of (D.26); (iii) follows from (D.25); and (iv) is valid when $A \geq 2$.

Case 2: when the sample size N is small. We now turn attention to the complement case where

$$
N < \widetilde{C} \frac{S\left(A+B\right) C^{\star}_{\mathsf{clipped}} \log\left(A+B\right)}{1-\gamma} \leq \frac{c_2 S\left(A+B\right) C^{\star}_{\mathsf{clipped}}}{\left(1-\gamma\right)^3 \varepsilon^2 \log\left(A+B\right)}.
$$

Given that the sample size is smaller than the one in Case 1, it is trivially seen that the minimax lower bound cannot be better than the former case, namely, we must have

$$
\inf_{(\widehat{\mu}, \widehat{\nu})} \sup_{\{\mathcal{MG}, \rho, d_{\mathsf{b}}\} \in \mathsf{MG}(C^{\star})} \mathbb{E}\left[V^{\star, \widehat{\nu}}\left(\rho\right) - V^{\widehat{\mu}, \star}\left(\rho\right)\right] \geq \varepsilon_0 = \sqrt{\frac{c_2}{2\widetilde{C}} \frac{1}{\left(1-\gamma\right) \log\left(A+B\right)}},
$$

where ε_0 is defined in (D.18).

Putting Case 1 and Case 2 together. Combining the above two cases reveals that: for any $\varepsilon \leq \varepsilon_0 = \sqrt{\frac{c_2}{2\widetilde{C}} \frac{1}{(1-\gamma)\log(A+B)}}$, one necessarily has

$$
\inf_{(\widehat{\mu}, \widehat{\nu})} \sup_{\{\mathcal{MG}, \rho, d_{\mathsf{b}}\} \in \mathsf{MG}(C^{\star})} \mathbb{E}\left[V^{\star, \widehat{\nu}}\left(\rho\right) - V^{\widehat{\mu}, \star}\left(\rho\right)\right] \geq \varepsilon
$$

if the sample size $N \leq \frac{c_2 S(A+B) C^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3 \varepsilon^2 \log(A+B)}$. This concludes the proof of Theorem 2.

E Auxiliary lemmas for Theorem 2

E.1 Proof of Lemma 6

First of all, it is straightforward to verify that: when initialized to any state $s \neq 0$, the MG will never leave the state s (by construction of P_θ). This combined with the fact that the rewards are zero whenever $s \neq 0$ gives

$$
V^{\mu,\nu}\left(s\right) = 0 \qquad \text{for all } s \neq 0. \tag{E.1}
$$

As a consequence, it suffices to compute $V_\theta^{\mu,\nu}(0)$. By virtue of the Bellman equation, we obtain

$$
V^{\mu,\nu}\left(0\right) = \mathop{\mathbb{E}}_{a\sim\mu(0), b\sim\nu(0)}\left[r\left(0,a,b\right) + \gamma \sum_{s'\in\mathcal{S}} P\left(s' \,|\, 0, a, b\right) V^{\mu,\nu}\left(s'\right)\right]
$$

$$\begin{aligned}
&\overset{\text{(i)}}{=} 1+\gamma \mathop{\mathbb{E}}_{a\sim\mu(0), b\sim\nu(0)}\left[P_\theta\left(0\,|\,0,a,b\right)V^{\mu,\nu}\left(0\right)\right] \\
&= 1+\gamma\sum_{a\in\mathcal{A}_{p,\theta}}\mu\left(a\,|\,0\right)\nu\left(0\,|\,0\right)P_\theta\left(0\,|\,0,a,0\right)V^{\mu,\nu}\left(0\right) \\
&\quad +\gamma\sum_{a\in\mathcal{A}_{p,\theta}}\mu\left(a\,|\,0\right)\sum_{b\neq 0}\nu\left(b\,|\,0\right)P_\theta\left(0\,|\,0,a,b\right)V^{\mu,\nu}\left(0\right) \\
&\quad +\gamma\sum_{a\in\mathcal{A}_{q,\theta}}\mu\left(a\,|\,0\right)\nu\left(0\,|\,0\right)P_\theta\left(0\,|\,0,a,0\right)V^{\mu,\nu}\left(0\right) \\
&\quad +\gamma\sum_{a\in\mathcal{A}_{q,\theta}}\mu\left(a\,|\,0\right)\sum_{b\neq 0}\nu\left(b\,|\,0\right)P_\theta\left(0\,|\,0,a,b\right)V^{\mu,\nu}\left(0\right) \\
&\overset{\text{(ii)}}{=} 1+\gamma\mu_{p,\theta}\nu_0 pV^{\mu,\nu}\left(0\right)+\gamma\mu_{p,\theta}\left(1-\nu_0\right)V^{\mu,\nu}\left(0\right) \\
&\quad +\gamma\left(1-\mu_{p,\theta}\right)\nu_0 qV^{\mu,\nu}\left(0\right)+\gamma\left(1-\mu_{p,\theta}\right)\left(1-\nu_0\right)V^{\mu,\nu}\left(0\right),
\end{aligned}$$

where we remind the readers of the quantities $\mu_{p,\theta} \coloneqq \sum_{a\in\mathcal{A}_{p,\theta}}\mu(a\,|\,0)$ and $\nu_0 \coloneqq \nu(0\,|\,0)$. Here, (i) makes use of the fact that $r(0,a,b)=1$ for all $(a,b)\in\mathcal{A}\times\mathcal{B}$ and (E.1), while (ii) follows from the definition of P_θ in (D.3). Rearrange terms to arrive at

$$\begin{aligned}
V^{\mu,\nu}\left(0\right) &= \frac{1}{1-\gamma\mu_{p,\theta}\nu_0 p-\gamma\mu_{p,\theta}\left(1-\nu_0\right)-\gamma\left(1-\mu_{p,\theta}\right)\nu_0 q-\gamma\left(1-\mu_{p,\theta}\right)\left(1-\nu_0\right)} \\
&= \frac{1}{1-\gamma+\gamma\mu_{p,\theta}\nu_0\left(1-p\right)+\gamma\left(1-\mu_{p,\theta}\right)\nu_0\left(1-q\right)}.
\end{aligned}$$

Next, let us rewrite $V^{\mu,\nu}(0)$ as follows

$$\begin{aligned}
V^{\mu,\nu}\left(0\right) &= \frac{1}{1-\gamma+\gamma\nu_0\left(1-p\right)-\gamma(1-\mu_{p,\theta})\nu_0\left(1-p\right)+\gamma\left(1-\mu_{p,\theta}\right)\nu_0\left(1-q\right)} \\
&= \frac{1}{1-\gamma+\gamma\nu_0\left(1-p\right)+\gamma\left(1-\mu_{p,\theta}\right)\nu_0\left(p-q\right)}.
\end{aligned}$$

Given that $p>q$, we see that: for any policy μ, the best response of the min-player would be to take $\nu_0=1$; and for any policy ν with $\nu_0\neq 0$, the best response of the max-player would be to set $\mu_{p,\theta}=1$. As a result, it is seen that (μ,ν) is an NE if and only if $\mu_{p,\theta}=1$ and $\nu_0=1$. This readily implies that the policy pair $(\mu^\star_\theta,\nu^\star_\theta)$ defined in (D.9) is an NE of $\mathcal{MG}_\theta$.

We are left with justifying that $\{\mathcal{MG}_\theta,\rho,d_{\mathsf{b}}\}\in\mathsf{MG}(C^\star_{\mathsf{clipped}})$. Towards this, we first note that: for any $a\in\mathcal{A}$, $b\in\mathcal{B}$, and any policy pair (μ,ν), we can invoke (D.6) to show that

$$\frac{\min\left\{d^{\mu,\nu}\left(0,a,b;\rho\right),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}\left(0,a,b\right)} \le \frac{\frac{1}{S(A+B)}}{\frac{1}{C^\star_{\mathsf{clipped}}S(A+B)}} = C^\star_{\mathsf{clipped}}, \tag{E.2}$$

and

$$\frac{\min\left\{d^{\mu,\nu}\left(1,a,b;\rho\right),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}\left(1,a,b\right)} \le \frac{\frac{1}{S(A+B)}}{\frac{1}{AB}\left(1-\frac{AB}{C^\star_{\mathsf{clipped}}S(A+B)}\right)} = \frac{1}{\frac{S(A+B)}{AB}-\frac{1}{C^\star_{\mathsf{clipped}}}} \le C^\star_{\mathsf{clipped}}, \tag{E.3}$$

where the last inequality holds as long as $C^\star_{\mathsf{clipped}}\ge\frac{2AB}{S(A+B)}$. In addition, for any $s\ge 2$, it is readily seen that $d^{\mu,\nu}(s,a,b;\rho)=0$, and therefore,

$$\frac{\min\left\{d^{\mu,\nu}\left(s,a,b;\rho\right),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}\left(s,a,b\right)} = 0. \tag{E.4}$$

Taking (E.2), (E.3) and (E.4) collectively gives

$$\max\left\{\sup_{\mu,s,a,b}\frac{\min\left\{d^{\mu,\nu^\star_\theta}\left(s,a,b;\rho\right),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}\left(s,a,b\right)},\sup_{\nu,s,a,b}\frac{\min\left\{d^{\mu^\star_\theta,\nu}\left(s,a,b;\rho\right),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}\left(s,a,b\right)}\right\}\le C^\star_{\mathsf{clipped}}. \tag{E.5}$$

We still need to justify that the inequality in (E.5) is tight. To do so, observe that for any $a \in \mathcal{A}_p$, one has

$$\begin{aligned}
d^{\mu_\theta^\star,\nu_\theta^\star}(0,a,0) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\left(s_t=0,a_t=a,b_t=0 \,|\, s_0\sim\rho;\mu_\theta^\star,\nu_\theta^\star\right)\\
&\overset{\text{(i)}}{=} (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\left(s_t=0 \,|\, s_0\sim\rho;\mu_\theta^\star,\nu_\theta^\star\right)\frac{1}{|\mathcal{A}_{p,\theta}|}\\
&\overset{\text{(ii)}}{\geq} \frac{1-\gamma}{|\mathcal{A}_{p,\theta}|}\sum_{t=0}^{\infty}\gamma^t p^t \overset{\text{(iii)}}{\geq} \frac{1-\gamma}{|\mathcal{A}_{p,\theta}|}\sum_{t=0}^{\infty}\gamma^{2t}\\
&= \frac{1-\gamma}{|\mathcal{A}_{p,\theta}|}\frac{1}{1-\gamma^2} = \frac{1}{|\mathcal{A}_{p,\theta}|}\frac{1}{1+\gamma} \geq \frac{1}{2|\mathcal{A}_{p,\theta}|}.
\end{aligned} \tag{E.6}$$

Here, (i) is valid since according to (D.9), $\mathbb{P}(a_t = a \,|\, s_t = 0; \mu_\theta^\star) = 1/|\mathcal{A}_{p,\theta}|$ for $a \in \mathcal{A}_{p,\theta}$ and $\mathbb{P}(b_t = 0 \,|\, s_t = 0; \nu_\theta^\star) = 1$; (ii) follows since

$$\mathbb{P}\left(s_t=0 \,|\, s_0\sim\rho;\mu_\theta^\star,\nu_\theta^\star\right) \geq \mathbb{P}\left(s_t=s_{t-1}=\cdots=s_0=0 \,|\, \mu_\theta^\star,\nu_\theta^\star\right) = p^t;$$

and (iii) relies on the fact that $p \geq \gamma$ (cf. (D.1)). Then we can invoke (E.6) and (D.6) to demonstrate that

$$\frac{\min\left\{d^{\mu_\theta^\star,\nu_\theta^\star}(0,a,0),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}(0,a,0)} = \frac{\frac{1}{S(A+B)}}{\frac{1}{C^\star_{\mathsf{clipped}}S(A+B)}} = C^\star_{\mathsf{clipped}} \qquad \text{for any } a\in\mathcal{A}_{p,\theta}. \tag{E.7}$$

Taking collectively (E.5) and (E.7) gives

$$\max\left\{\sup_{\mu,s,a,b}\frac{\min\left\{d^{\mu,\nu_\theta^\star}(s,a,b;\rho),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}(s,a,b)},\ \sup_{\nu,s,a,b}\frac{\min\left\{d^{\mu_\theta^\star,\nu}(s,a,b;\rho),\frac{1}{S(A+B)}\right\}}{d_{\mathsf{b}}(s,a,b)}\right\} = C^\star_{\mathsf{clipped}}.$$

This allows one to conclude that $\{\mathcal{MG}_\theta, \rho, d_{\mathsf{b}}\} \in \mathsf{MG}(C^\star_{\mathsf{clipped}})$.

E.2 Proof of Lemma 7

Recall that $M_a \sim \mathsf{Binomial}(N, d_{\mathsf{b}}(0,a,0))$. Invoke the Chernoff bound (e.g., Vershynin (2018, Exercise 2.3.5)) to show the existence of some universal constant $c_{\mathsf{ch}} > 0$ such that

$$\mathbb{P}\Big(\left|M_a - Nd_{\mathsf{b}}(0,a,0)\right| \geq \delta N d_{\mathsf{b}}(0,a,0)\Big) \leq 2\exp\left(-c_{\mathsf{ch}}Nd_{\mathsf{b}}(0,a,0)\,\delta^2\right) \tag{E.8}$$

holds for any $\delta \in (0,1]$. For $\widetilde{C}$ sufficiently large, one has (see (D.16))

$$N \geq \frac{4}{c_{\mathsf{ch}}}C^\star_{\mathsf{clipped}}S(A+B)\log A,$$

or equivalently,

$$Nd_{\mathsf{b}}(0,a,0) \geq \frac{4}{c_{\mathsf{ch}}}\log A.$$

Take

$$\delta := \sqrt{\frac{4\log A}{c_{\mathsf{ch}}Nd_{\mathsf{b}}(0,a,0)}} \leq 1$$

in (E.8) to show that with probability exceeding $1 - 2A^{-4}$,

$$M_a \leq Nd_{\mathsf{b}}(0,a,0) + \sqrt{Nd_{\mathsf{b}}(0,a,0)}\sqrt{\frac{4\log A}{c_{\mathsf{ch}}}} \overset{\text{(i)}}{\leq} \frac{c_2}{(1-\gamma)^3\varepsilon^2\log A} + \sqrt{\frac{8c_2}{c_{\mathsf{ch}}(1-\gamma)^3\varepsilon^2}}$$

$$\overset{\text{(ii)}}{\leq} \frac{c_2 + \sqrt{8c_2/c_{\mathsf{ch}}}}{(1-\gamma)^3 \varepsilon^2 \log A} = \frac{c_4}{(1-\gamma)^3 \varepsilon^2 \log A} \tag{E.9}$$

holds for some universal constant $c_4 = c_2 + \sqrt{8c_2/c_{\mathsf{ch}}}$, where (i) makes use of Condition (D.17), and (ii) is valid as long as $\varepsilon \leq \frac{1}{(1-\gamma)\log A}$. Moreover, we can also invoke (E.8) and (D.16) to show that with probability exceeding $1 - 2A^{-4}$,

$$M_a \geq N d_{\mathsf{b}}(0,a,0) - \sqrt{N d_{\mathsf{b}}(0,a,0)} \sqrt{\frac{2\log A}{c_{\mathsf{ch}}}} \overset{\text{(iii)}}{\geq} \widetilde{C}\frac{\log A}{1-\gamma} - \sqrt{\frac{2\widetilde{C}}{c_{\mathsf{ch}}}\frac{\log^2 A}{1-\gamma}} \overset{\text{(iv)}}{\geq} \frac{c_3 \log A}{1-\gamma} \tag{E.10}$$

for some universal constant $c_3 = \widetilde{C}/2$. Here, (iii) holds since the function $x^2 - \sqrt{\frac{2\log A}{c_{\mathsf{ch}}}}x$ is monotonically increasing when $x \geq \sqrt{\frac{\log A}{c_{\mathsf{ch}}}}$, and

$$N d_{\mathsf{b}}(0,a,0) \geq \widetilde{C}\frac{\log A}{1-\gamma} \geq \sqrt{\frac{\log A}{c_{\mathsf{ch}}}}$$

when $\widetilde{C} > 0$ is sufficiently large; (iv) holds for $\widetilde{C} > 0$ sufficiently large. Taking (E.9) and (E.10) collectively gives

$$\mathbb{P}\left(\frac{c_3 \log A}{1-\gamma} \leq M_a \leq \frac{c_4}{(1-\gamma)^3 \varepsilon^2 \log A}\right) \geq 1 - \frac{2}{A^4}.$$

E.3 Proof of Lemma 8

Given that $B_p \sim \mathsf{Binomial}(M,p)$, we have $M - B_p \sim \mathsf{Binomial}(M, 1-p)$. In view of the Chernoff bound (cf. Vershynin (2018, Exercise 2.3.5)), we know that for any $\delta \in (0,1]$

$$\begin{aligned} \mathbb{P}\Big(|B_p - Mp| > \delta M(1-p)\Big) &= \mathbb{P}\Big(|M - B_p - M(1-p)| > \delta M(1-p)\Big) \\ &\leq 2\exp\left(-c_{\mathsf{ch}} M(1-p)\delta^2\right) \end{aligned} \tag{E.11}$$

for some universal constant $c_{\mathsf{ch}} > 0$. Recall that $M \geq \frac{c_3 \log A}{1-\gamma}$. By taking

$$\delta := \sqrt{\frac{4\log A}{c_{\mathsf{ch}} M(1-p)}}$$

we can guarantee that

$$\delta \leq \sqrt{\frac{4(1-\gamma)}{c_{\mathsf{ch}} c_3 (1-p)}} = \sqrt{\frac{4(1-\gamma)}{c_{\mathsf{ch}} c_3 \left(1 - \gamma - 14\frac{(1-\gamma)^2 \varepsilon}{\gamma}\right)}} \leq \sqrt{\frac{4}{c_{\mathsf{ch}} c_3 \left(1 - 14\frac{(1-\gamma)\varepsilon}{\gamma}\right)}} \overset{\text{(i)}}{\leq} \sqrt{\frac{8}{c_{\mathsf{ch}} c_3}} \overset{\text{(ii)}}{\leq} 1.$$

Here, (i) holds when $\varepsilon \leq \frac{1}{42(1-\gamma)}$ and $\gamma \geq 2/3$, while (ii) is valid as long as $c_3 = \widetilde{C}/2 \geq 8/c_{\mathsf{ch}}$ (a condition that can be guaranteed as long as $\widetilde{C}$ is sufficiently large). Then the inequality (E.11) tells us that

$$\mathbb{P}\left(|B_p - Mp| > \sqrt{\frac{4\log A}{c_{\mathsf{ch}}}M(1-p)}\right) \leq \frac{2}{A^4}.$$

A similar argument leads to

$$\mathbb{P}\left(|B_q - Mq| \geq \sqrt{\frac{4\log A}{c_{\mathsf{ch}}}M(1-q)}\right) \leq \frac{2}{A^4}.$$

In view of these concentration bounds, we can take

$$E_M \coloneqq \left\{0 \le n \le M : \lfloor \min\left\{Mp - \Delta_p, Mq - \Delta_q\right\} \rfloor \le n \le \lceil \max\left\{Mp + \Delta_p, Mq + \Delta_q\right\} \rceil \right\},$$

where we define, for every $x \in (0,1)$,

$$\Delta_x \coloneqq \sqrt{\frac{4\log A}{c_{\mathsf{ch}}} M\left(1-x\right)}.$$

It is clear from the above argument that (D.21) is guaranteed to hold.

We are left with checking (D.22). Towards this, we first make the observation that

$$\begin{aligned}\frac{\mathbb{P}\left(B_q = m\right)}{\mathbb{P}\left(B_p = m\right)} &= \frac{q^m\left(1-q\right)^{M-m}}{p^m\left(1-p\right)^{M-m}} = \exp\left(-m\log\frac{p}{q} - \left(M-m\right)\log\frac{1-p}{1-q}\right)\\ &= \exp\left(-M\left[\mathsf{KL}\left(p\,\|\,q\right) + \left(\frac{m}{M}-p\right)\log\frac{p\left(1-q\right)}{q\left(1-p\right)}\right]\right),\end{aligned}$$

where $\mathsf{KL}(p\,\|\,q)$ denotes the Kullback-Leibler (KL) divergence between $\mathsf{Bernoulli}(p)$ and $\mathsf{Bernoulli}(q)$ (Tsybakov and Zaiats, 2009):

$$\mathsf{KL}\left(p\,\|\,q\right) \coloneqq p\log\frac{p}{q} + \left(1-p\right)\log\frac{1-p}{1-q}.$$

Moreover, it is seen that

$$\mathsf{KL}\left(p\,\|\,q\right) \overset{\text{(i)}}{\le} \frac{\left(p-q\right)^2}{p\left(1-p\right)} = \frac{784\left(1-\gamma\right)^4\varepsilon^2}{p\left(1-\gamma-14\frac{(1-\gamma)^2\varepsilon}{\gamma}\right)\gamma^2} \overset{\text{(ii)}}{\le} \frac{3528\left(1-\gamma\right)^3\varepsilon^2}{1-14\frac{(1-\gamma)\varepsilon}{\gamma}} \le 7056\left(1-\gamma\right)^3\varepsilon^2, \tag{E.12}$$

where (i) follows from Li et al. (2022c, Lemma 10) and (ii) relies on (D.2). In addition, for any $m \in E_M$, we have $m \le Mp + \Delta_q$, which in turn reveals that

$$\frac{m}{M} - p \le \frac{\Delta_q}{M} \le \sqrt{\frac{4\log A}{c_{\mathsf{ch}}M}\left(1-\gamma+14\frac{\left(1-\gamma\right)^2\varepsilon}{\gamma}\right)} \le \sqrt{\frac{6\left(1-\gamma\right)\log A}{c_{\mathsf{ch}}M}} \tag{E.13}$$

provided that $\varepsilon \le \frac{1}{42(1-\gamma)}$ and $\gamma \le 2/3$. Recalling that $p > q$, we can also see that

$$\begin{aligned}0 < \log\frac{p\left(1-q\right)}{q\left(1-p\right)} &= \log\left(1+\frac{p-q}{q\left(1-p\right)}\right) \overset{\text{(i)}}{\le} \frac{p-q}{q\left(1-p\right)} \overset{\text{(ii)}}{\le} \frac{28\left(1-\gamma\right)^2\varepsilon}{\gamma\left(1-\gamma-14\frac{(1-\gamma)^2\varepsilon}{\gamma}\right)}\\ &\le \frac{42\left(1-\gamma\right)\varepsilon}{1-14\frac{(1-\gamma)\varepsilon}{\gamma}} \le \frac{42\left(1-\gamma\right)\varepsilon}{1-\frac{1}{3\gamma}} \le 84\left(1-\gamma\right)\varepsilon,\end{aligned} \tag{E.14}$$

where (i) exploits the elementary inequality $\log(1+x) \le x$ for all $x \ge 0$, and (ii) follows from (D.2). Taking (E.12), (E.13) and (E.14) collectively yields

$$\begin{aligned}\frac{\mathbb{P}\left(B_q = m\right)}{\mathbb{P}\left(B_p = m\right)} &\ge \exp\left[-M\left(7056\left(1-\gamma\right)^3\varepsilon^2 + 84\left(1-\gamma\right)\varepsilon\sqrt{\frac{6\left(1-\gamma\right)\log A}{c_{\mathsf{ch}}M}}\right)\right]\\ &\overset{\text{(i)}}{\ge} \exp\left[-\left(\frac{7056c_4}{\log A} + 84\sqrt{\frac{6c_4}{c_{\mathsf{ch}}}}\right)\right] \overset{\text{(ii)}}{\ge} \exp\left[-\left(\frac{7056c_4}{\log 2} + 84\sqrt{\frac{6c_4}{c_{\mathsf{ch}}}}\right)\right] \overset{\text{(iii)}}{\ge} \frac{1}{2}.\end{aligned}$$

Here, (i) follows from the condition that

$$M \le \frac{c_4}{\left(1-\gamma\right)^3\varepsilon^2\log A},$$

(ii) holds when $A \ge 2$, and (iii) holds with the proviso that $c_4 = 2c_2 + \sqrt{8c_2/c_{\mathsf{ch}}}$ is sufficiently small, which can happen as long as c_2 is sufficiently small.

References

Abe, K. and Kaneko, Y. (2020). Off-policy exploitability-evaluation in two-player zero-sum markov games. *arXiv preprint arXiv:2007.02141*.

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

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

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR.

Bai, Y. and Jin, C. (2020). Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR.

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

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., and Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

Bertsekas, D. P. (2017). *Dynamic programming and optimal control (4th edition)*. Athena Scientific.

Brown, N. and Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890.

Cen, S., Wei, Y., and Chi, Y. (2021). Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34.

Chang, J. D., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. (2021). Mitigating covariate shift in imitation learning via offline data without great coverage. *arXiv preprint arXiv:2106.03207*.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. (2021a). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

Chen, Z., Zhou, D., and Gu, Q. (2021b). Almost optimal algorithms for two-player markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*.

Ciesielski, K. (2007). On stefan banach and some of his results. *Banach Journal of Mathematical Analysis*, 1(1):1–10.

Cui, Q. and Du, S. S. (2022a). Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *arXiv preprint arXiv:2206.00177*.

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

Cui, Q. and Yang, L. F. (2021). Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504. PMLR.

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

Daskalakis, C., Foster, D. J., and Golowich, N. (2020). Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540.

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

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

Freund, Y. and Schapire, R. E. (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103.

Hansen, T. D., Miltersen, P. B., and Zwick, U. (2013). Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16.

Hu, J. and Wellman, M. P. (2003). Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., and Ruderman, A. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865.

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

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

Jin, Y., Muthukumar, V., and Sidford, A. (2022). The complexity of infinite-horizon general-sum stochastic games. *arXiv preprint arXiv:2204.04186*.

Jin, Y. and Sidford, A. (2021). Towards tight bounds on the sample complexity of average-reward MDPs. In *International Conference on Machine Learning*, pages 5055–5064. PMLR.

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

Johnstone, I. M. (2017). Gaussian estimation: Sequence and wavelet models. *Unpublished manuscript*.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc.

Lagoudakis, M. G. and Parr, R. (2002). Value function approximation in zero-sum markov games. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 283–292.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Li, G., Chi, Y., Wei, Y., and Chen, Y. (2022a). Minimax-optimal multi-agent RL in zero-sum Markov games with a generative model. *in preparation*.

Li, G., Ma, C., and Srebro, N. (2022b). Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *arXiv preprint arXiv:2205.10671*.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022c). Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*.

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

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

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

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

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

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

Lu, M., Min, Y., Wang, Z., and Yang, Z. (2022). Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. *arXiv preprint arXiv:2205.13589*.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3).

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

Munos, R. (2003). Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567.

Munos, R. (2007). Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561.

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

Perolat, J., Scherrer, B., Piot, B., and Pietquin, O. (2015). Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*, pages 1321–1329. PMLR.

Raghavan, T. (1994). Zero-sum two-person games. *Handbook of game theory with economic applications*, 2:735–768.

Rakhlin, S. and Sridharan, K. (2013). Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Neural Information Processing Systems (NeurIPS)*.

Roughgarden, T. (2016). *Twenty lectures on algorithmic game theory*. Cambridge University Press.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Shao, J. (2003). *Mathematical Statistics*. Springer Texts in Statistics. Springer.

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

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. *accepted to International Conference on Machine Learning (ICML)*.

Sidford, A., Wang, M., Yang, L., and Ye, Y. (2020). Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR.

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

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

Tsybakov, A. B. and Zaiats, V. (2009). *Introduction to nonparametric estimation*, volume 11. Springer.

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

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

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

Wang, B., Yan, Y., and Fan, J. (2021). Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in Neural Information Processing Systems*, 34.

Wang, X., Cui, Q., and Du, S. S. (2022). On gap-dependent bounds for offline reinforcement learning. *arXiv preprint arXiv:2206.00177*.

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

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

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. (2020). Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *arXiv preprint arXiv:2106.04895*.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. (2022). Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*.

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

Yang, Y. and Wang, J. (2020). An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*.

Yin, M., Bai, Y., and Wang, Y.-X. (2021a). Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*.

Yin, M., Bai, Y., and Wang, Y.-X. (2021b). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. (2021c). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representations*.

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

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., and Finn, C. (2021a). Conservative data sharing for multi-task offline reinforcement learning. *arXiv preprint arXiv:2109.08128*.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. (2021b). Combo: Conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. (2020). MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*.

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

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

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar, T. (2021b). Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12):5925–5940.

Zhang, X., Chen, Y., Zhu, J., and Sun, W. (2021c). Corruption-robust offline reinforcement learning. *arXiv preprint arXiv:2106.06630*.

Zhang, Z., Zhou, Y., and Ji, X. (2020b). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33.

Zhao, Y., Tian, Y., Lee, J. D., and Du, S. S. (2021). Provably efficient policy gradient methods for two-player zero-sum Markov games. *arXiv preprint arXiv:2102.08903*.

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