Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Model-Based Reinforcement Learning for Offline Zero-Sum Markov Games

Yuling Yan, Gen Li, Yuxin Chen, Jianqing Fan

To cite this article:
Yuling Yan, Gen Li, Yuxin Chen, Jianqing Fan (2024) Model-Based Reinforcement Learning for Offline Zero-Sum Markov Games. Operations Research 72(6):2430-2445. https://doi.org/10.1287/opre.2022.0342

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 72, No. 6, November–December 2024, pp. 2430–2445
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Model-Based Reinforcement Learning for Offline Zero-Sum Markov Games

Yuling Yan,[a] Gen Li,[b] Yuxin Chen,[c,*] Jianqing Fan[d]

[a] Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [b] Department of Statistics, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China; [c] Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [d] Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544

*Corresponding author

Contact: yulingy@mit.edu, https://orcid.org/0000-0002-5747-8182 (YY); genli@cuhk.edu.hk (GL); yuxinc@wharton.upenn.edu, https://orcid.org/0000-0001-9256-5815 (YC); jqfan@princeton.edu, https://orcid.org/0000-0003-3250-7677 (JF)

Received: June 30, 2022
Revised: August 3, 2023
Accepted: February 13, 2024
Published Online in Articles in Advance: April 2, 2024

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2022.0342

Abstract. This paper makes progress toward 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, in which 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 the value iteration with lower confidence bounds for zero-sum Markov games—that provably finds an ε-approximate Nash equilibrium with a sample complexity no larger than $\frac{C^\star_{\text{clipped}} S(A+B)}{(1-\gamma)^3\varepsilon^2}$ (up to some log factor). Here, $C^\star_{\text{clipped}}$ 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 $\left(0, \frac{1}{1-\gamma}\right]$. Our sample complexity bound strengthens prior art by a factor of $\min\{A, B\}$, achieving minimax optimality for a broad regime of interest. An appealing feature of our result lies in its algorithmic simplicity, which reveals the unnecessity of variance reduction and sample splitting in achieving sample optimality.

Funding: Y. Yan is supported in part by the Charlotte Elizabeth Procter Honorific Fellowship from Princeton University and the Norbert Wiener Postdoctoral Fellowship from MIT. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the Google Research Scholar Award, the Air Force Office of Scientific Research [Grant FA9550-22-1-0198], the Office of Naval Research [Grant N00014-22-1-2354], and the National Science Foundation [Grants CCF-2221009, CCF-1907661, IIS-2218713, DMS-2014279, and IIS-2218773]. J. Fan is supported in part by the National Science Foundation [Grants DMS-1712591, DMS-2052926, DMS-2053832, and DMS-2210833] and Office of Naval Research [Grant N00014-22-1-2340].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2022.0342.

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

1. Introduction

Multiagent 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 (Shalev-Shwartz et al. 2016, Baker et al. 2019, Berner et al. 2019, Brown and Sandholm 2019, Jaderberg et al. 2019, 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 are 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. Whereas computational intractability has been observed when calculating NEs in general-sum MGs and/or MGs with more than two players (Daskalakis et al. 2009, Daskalakis 2013), 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, Xie et al. 2020, Zhang et al. 2020a, Liu et al. 2021, Tian et al. 2021). 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. 2022b); 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 Jin et al. (2021a), Song et al. (2021), Mao and Başar (2022), and Daskalakis et al. (2023) for similar accomplishments in learning coarse correlated equilibria in multiplayer 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—for example, 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 data set of logged interactions.

1.1. 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 one hand, if the available data set 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—that is, 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 define shortly in Assumption 1), Cui and Du (2022b) demonstrate how to find ε-Nash solutions in a finite-horizon, two-player, zero-sum MG once the number of sample rollouts exceeds

$$\tilde{O}\left(\frac{C^\star H^3 SAB}{\varepsilon^2}\right). \tag{1}$$

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 $\tilde{O}(\cdot)$ denotes the order-wise scaling with all logarithmic dependency hidden.

Despite being an intriguing polynomial sample complexity bound, a shortfall of (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 in a two-player, zero-sum Markov game—and, if so, how to accomplish it—is the key question to be investigated in the current paper.

1.2. 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 the value iteration with lower confidence bounds for zero-sum Markov game (VI-LCB-Game)—that is capable of learning an ε-approximate Nash equilibrium with sample complexity

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

where $C^\star_{\text{clipped}}$ is the so-called clipped unilateral concentrability coefficient (formalized in Assumption 2) and always satisfies $C^\star_{\text{clipped}} \le C^\star$. 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 equivalence). To demonstrate that this bound is essentially unimprovable, 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 such as reference-advantage decomposition. The fact that our sample complexity result holds for the full ε-range (i.e., any $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$) unveils that sample efficiency is achieved without incurring any burn-in cost.

Finally, when 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 in a two-player, zero-sum Markov game, on which we elaborate toward the end of Section 3.

1.3. Notation

Before proceeding, let us introduce some notation that is used throughout. With slight abuse of notation, we 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 entry-wise fashion. For a finite set $\mathcal{S} = \{1, \ldots, S\}$, we let $\Delta(\mathcal{S}) := \{x \in \mathbb{R}^S \,|\, 1^\top x = 1, x \geq 0\}$ 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 data set.

2.1. Preliminaries

2.1.1. Zero-Sum, Two-Player Markov Games. Consider a discounted, infinite-horizon, zero-sum MG (Shapley 1953, Littman 1994) as represented by the tuple $\mathcal{MG} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, \gamma)$. Here, $\mathcal{S} = \{1, \ldots, S\}$ is the shared state space; $\mathcal{A} = \{1, \ldots, A\}$ (respectively, $\mathcal{B} = \{1, \ldots, B\}$) is the action space of the max-player (min-player); $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ is the (a priori unknown) probability transition kernel, where $P(s'|s,a,b)$ denotes the probability of transitioning 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 in which S, A, B, and $\frac{1}{1-\gamma}$ could all be large. Additionally, for notational simplicity, we define the vector $P_{s,a,b} \in \mathbb{R}^{1 \times S}$ as $P_{s,a,b} := P(\cdot|s,a,b)$ for any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$.

2.1.2. 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, $\mu(\cdot|s) \in \Delta(\mathcal{A})$ ($\nu(\cdot|s) \in \Delta(\mathcal{B})$) specifies the action selection probability of the max-player (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

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

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 $a_t \sim \mu(\cdot|s_t)$ and $b_t \sim \nu(\cdot|s_t)$ independently conditional on the past) and the probability transition kernel P (i.e., $s_{t+1} \sim P(\cdot|s_t,a_t,b_t)$ 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:

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

where the actions are drawn from $\mu \times \nu$ except for the initial time step (namely, for any $t \geq 1$, we execute $a_t \sim \mu(\cdot|s_t)$ and $b_t \sim \nu(\cdot|s_t)$ 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):

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

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:

$$d^{\mu,\nu}(s;\rho) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s | s_0 \sim \rho; \mu, \nu), \qquad \forall s \in \mathcal{S}, \tag{2}$$

$$d^{\mu,\nu}(s,a,b;\rho) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a, b_t = b | s_0 \sim \rho; \mu, \nu), \qquad \forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}, \tag{3}$$

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 preceding definition that

$$d^{\mu,\nu}(s,a,b;\rho) = d^{\mu,\nu}(s;\rho)\mu(a|s)\nu(b|s). \tag{4}$$

2.1.3. 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 an 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 the player's 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

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

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:

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

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 $s \in \mathcal{S}$,

$$V^{\mu,\star}(s) := \min_\nu V^{\mu,\nu}(s) \quad \text{and} \quad V^{\star,\nu} := \max_\mu V^{\mu,\nu}(s),$$

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:

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

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

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

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

2.2. Offline Data Set (Batch Data Set)

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

$$(s_i, a_i, b_i) \overset{\text{i.i.d.}}{\sim} d_b \quad \text{and} \quad s_i' \overset{\text{ind.}}{\sim} P(\cdot | s_i, a_i, b_i). \tag{5}$$

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

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) introduce the following unilateral concentrability condition.

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

$$C^\star := \max\left\{ \sup_{\mu,s,a,b} \frac{d^{\mu,\nu^\star}(s,a,b;\rho)}{d_b(s,a,b)}, \sup_{\nu,s,a,b} \frac{d^{\mu^\star,\nu}(s,a,b;\rho)}{d_b(s,a,b)} \right\} \tag{6}$$

is 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 $\frac{d^{\mu,\nu}(s,a,b;\rho)}{d_b(s,a,b)}$ over all (μ, ν) simultaneously) as (6) freezes the policy of one side, 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 data set to cover the part of the state–action space reachable by a given policy pair $(\mu^\star, \nu^\star)$) as (6) 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 in which single-policy concentrability suffices for finding the optimal policy (Rashidinejad et al. 2021, Xie et al. 2021, Li et al. 2024b).

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*

$$C^\star_{\text{clipped}} := \max\left\{ \sup_{\mu,s,a,b} \frac{\min\left\{ d^{\mu,\nu^\star}(s,a,b;\rho), \frac{1}{S(A+B)} \right\}}{d_b(s,a,b)}, \sup_{\nu,s,a,b} \frac{\min\left\{ d^{\mu^\star,\nu}(s,a,b;\rho), \frac{1}{S(A+B)} \right\}}{d_b(s,a,b)} \right\} \tag{7}$$

is finite, where we define $0/0 = 0$ *by convention. This quantity* $C^\star_{\text{clipped}}$ *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 a (slight) relaxation of Assumption 1. Comparing (7) with (6) immediately reveals that

$$C^\star \ge C^\star_{\text{clipped}}$$

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

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

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

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

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

$$\hat{P}(s'|s,a,b)=\begin{cases}\dfrac{1}{N(s,a,b)}\displaystyle\sum_{i=1}^{N}\mathbb{1}\{(s_i,a_i,b_i,s_i')=(s,a,b,s')\}, & \text{if } N(s,a,b)>0\\ \dfrac{1}{S}, & \text{if } N(s,a,b)=0\end{cases}\tag{8}$$

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

$$\hat{r}(s,a,b)=\begin{cases}r(s,a,b), & \text{if } N(s,a,b)>0\\ 0, & \text{if } N(s,a,b)=0\end{cases}\tag{9}$$

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 $\widehat{\mathcal{MG}}=(\mathcal{S},\mathcal{A},\mathcal{B},\hat{P},\hat{r},\gamma)$.

3.1.2. Pessimistic Bellman Operators. Recall that the classic Bellman operator $\mathcal{T}:\mathbb{R}^{SAB}\to\mathbb{R}^{SAB}$ is defined such that (Shapley 1953, Lagoudakis and Parr 2002), 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, that is,

$$V(s):=\max_{\mu_s\in\Delta(\mathcal{A})}\min_{\nu_s\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim\mu_s,b\sim\nu_s}[Q(s,a,b)],\quad\forall s\in\mathcal{S}.\tag{10}$$

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 $\hat{\mathcal{T}}^-_{\text{pe}}$ ($\hat{\mathcal{T}}^+_{\text{pe}}$) for the max-player (min-player) as follows: for every $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$,

$$\hat{\mathcal{T}}^-_{\text{pe}}(Q)(s,a,b):=\max\{\hat{r}(s,a,b)+\gamma\hat{P}_{s,a,b}V-\beta(s,a,b;V),\,0\},\tag{11a}$$

$$\hat{\mathcal{T}}^+_{\text{pe}}(Q)(s,a,b):=\min\left\{\hat{r}(s,a,b)+\gamma\hat{P}_{s,a,b}V+\beta(s,a,b;V),\,\frac{1}{1-\gamma}\right\},\tag{11b}$$

where V is again defined in (10). 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 $\hat{\mathcal{T}}^-_{\text{pe}}$ ($\hat{\mathcal{T}}^+_{\text{pe}}$) produce a conservative estimate of the Q-function from the max-player's (min-player's) viewpoint. Here and throughout, we choose this term based on Bernstein-style concentration bounds; specifically, we take

$$\beta(s,a,b;V)=\min\left\{\max\left\{\sqrt{\frac{C_{\text{b}}\log\frac{N}{\delta}}{N(s,a,b)}\mathsf{Var}_{\hat{P}_{s,a,b}}(V)},\ \frac{2C_{\text{b}}\log\frac{N}{\delta}}{(1-\gamma)N(s,a,b)}\right\},\frac{1}{1-\gamma}\right\}+\frac{4}{N}\tag{12}$$

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

$$\mathsf{Var}_{\hat{P}_{s,a,b}}(V):=\hat{P}_{s,a,b}V^2-(\hat{P}_{s,a,b}V)^2.\tag{13}$$

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

$$\|\hat{\mathcal{T}}^-_{\text{pe}}(Q_1)-\hat{\mathcal{T}}^-_{\text{pe}}(Q_2)\|_\infty\leq\gamma\|Q_1-Q_2\|_\infty\quad\text{and}$$
$$\|\hat{\mathcal{T}}^+_{\text{pe}}(Q_1)-\hat{\mathcal{T}}^+_{\text{pe}}(Q_2)\|_\infty\leq\gamma\|Q_1-Q_2\|_\infty;\tag{14}$$

see Lemma 1 for precise statements.

3.1.3. 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 $\{Q^-_{\text{pe},t}\}$, the policy iterates $\{\mu^-_t\}$ and $\{\nu^-_t\}$, and the value function iterates $\{V^-_{\text{pe},t}\}$ from the max-player's perspective; at the same time, it also maintains an analogous group of iterates $\{Q^+_{\text{pe},t}\}$, $\{\mu^+_t\}$ and $\{\nu^+_t\}$, and $\{V^+_{\text{pe},t}\}$ 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 write $\mu(s):=\mu(\cdot|s)\in\Delta(\mathcal{A})$ and $\nu(s):=\nu(\cdot|s)\in\Delta(\mathcal{B})$

whenever it is clear from the context. In each round $t = 1, 2, \ldots$, we carry out the following update rules:

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

$$Q^-_{\mathsf{pe},t} = \hat{\mathcal{T}}^-_{\mathsf{pe}}(Q^-_{\mathsf{pe},t-1}). \tag{15}$$

The γ-contraction property (14) 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 (15). Specifically, for each $s \in \mathcal{S}$, we compute the Nash equilibrium $(\mu^-_t(s), \nu^-_t(s)) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$ 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 include linear programming and no-regret learning (Raghavan 1994, Freund and Schapire 1999, 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 $(\mu^-_t(s), \nu^-_t(s))$ as follows:

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

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

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

$$Q^-_{\mathsf{pe}} := Q^-_{\mathsf{pe},T} \quad \text{and} \quad Q^+_{\mathsf{pe}} := Q^+_{\mathsf{pe},T}, \tag{16}$$

in addition to two sets of policy estimates

$$(\mu^-, \nu^-) := (\mu^-_T, \nu^-_T) \quad \text{and} \quad (\mu^+, \nu^+) := (\mu^+_T, \nu^+_T). \tag{17}$$

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

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

The full algorithm is summarized in Algorithm 1.

Algorithm 1 (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 $\hat{P}$ as (8) and the empirical reward function $\hat{r}$ as (9).

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

- Update

$$\begin{aligned} Q^-_{\mathsf{pe},t}(s,a,b) &= \hat{\mathcal{T}}^-_{\mathsf{pe}}(Q^-_{\mathsf{pe},t-1}) \\ &= \max\{\hat{r}(s,a,b) + \gamma \hat{P}_{s,a,b} V^-_{\mathsf{pe},t-1} \\ &\quad - \beta(s,a,b; V^-_{\mathsf{pe},t-1}),\ 0\}, \\ Q^+_{\mathsf{pe},t}(s,a,b) &= \hat{\mathcal{T}}^+_{\mathsf{pe}}(Q^+_{\mathsf{pe},t-1}) \\ &= \min\Big\{\hat{r}(s,a,b) + \gamma \hat{P}_{s,a,b} V^+_{\mathsf{pe},t-1} \\ &\quad + \beta(s,a,b; V^+_{\mathsf{pe},t-1}),\ \frac{1}{1-\gamma}\Big\}, \end{aligned}$$

where

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

for some sufficiently large constant $C_{\mathsf{b}} > 0$ with $\mathsf{Var}_{\hat{P}_{s,a,b}}(V)$ defined in (13).

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

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

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

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

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

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

3.2. Theoretical Guarantees

Our main result is to uncover the intriguing sample efficiency of the proposed model-based algorithm. This is formally stated as follows with the proof postponed to Section 6.

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

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

as long as the sample size exceeds

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

for some sufficiently large constant $c_1 > 0$.

Remark 1. Our result and analysis are inspired by prior works that show that model-based RL achieves,

in multiple settings, sample efficiency without the need of variance reduction (Agarwal et al. 2020; Li et al. 2024a, b). The proof of this sample complexity bound entails several key analysis ingredients: (i) a leave-one-out analysis argument that proves effective in decoupling complicated statistical dependency and (ii) a careful self-bounding trick (i.e., upper bounding a certain quantity by a contraction of itself in addition to some other error terms) to derive a sharp control of the target duality gap. See Section 6 for details. Although techniques such as leave-one-out analysis are used in some prior RL literature (Agarwal et al. 2020; Li et al. 2024a, b), as far as we know, our work applies this technique for the first time to multiagent reinforcement learning. It has been observed that extending the algorithmic or analysis ideas in single-agent RL to the multiagent counterpart often leads to suboptimal sample complexity bounds that scale linearly in the total number of joint actions AB (Zhang et al. 2020a, Cui and Du 2022b). In contrast, our analysis framework leads to an optimal sample complexity bound that scales linearly in the total number of individual actions $A + B$.

Remark 2. A line of recent works focuses on instance-optimality of RL algorithms (Khamaru et al. 2021a, b; Mou et al. 2022). However, it remains challenging to establish instance-dependent bounds for multiagent RL even in two-player, zero-sum Markov games because of the difficulties arising from offline data and multiagent settings. Unlike RL with a generative model (simulator) that can generate independent samples for all state–action pairs, offline RL suffers from substantially more challenges, such as distribution shift and limited data coverage, making it more difficult to derive instance-dependent error bounds. In addition, the prior literature Khamaru et al. (2021a) that establishes instance optimality of variance-reduced Q-learning algorithms for the optimal value estimation problem requires one of the following two conditions: the optimal policy is unique or a meaningful sample complexity bound that depends on an optimality gap can be obtained. However, neither condition has a direct analog in zero-sum Markov games; this is because the Nash equilibrium in a zero-sum Markov game is not unique in general, and there is no well-defined analog of optimality gap for zero-sum Markov games. Detailed discussion on the challenges and difficulty of extending our analysis to develop instance-dependent error bounds can be found in Section 6.5.

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

$$\tilde{O}\left(\frac{C^{\star}_{\mathsf{clipped}} S(A+B)}{(1-\gamma)^3 \varepsilon^2}\right), \tag{18}$$

which accommodates any target accuracy within the range $\left(0, \frac{1}{1-\gamma}\right]$. In addition to linear dependency on $C^{\star}_{\mathsf{clipped}}$, the sample complexity bound (18) 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 such as variance reduction (Zhang et al. 2020a, b; Li et al. 2021; Xie et al. 2021; Yan et al. 2023).

As it turns out, the preceding sample complexity theory for Algorithm 1 matches the minimax lower limit modulo some logarithmic term as asserted by the following theorem. This minimax lower bound—whose proof is postponed to Online Appendix EC.2—is inspired by prior lower bound theory for single-agent MDPs (e.g., Azar et al. 2013, Li et al. 2024b) and might shed light on how to establish lower bounds for other game-theoretic settings.

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

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

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

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

Here, the infimum is taken over all estimators $(\hat{\mu}, \hat{\nu})$ for the Nash equilibrium based on the batch data set $\mathcal{D} = \{(s_i, a_i, b_i, s_i')\}_{i=1}^{n}$ generated according to (5).

Remark 3. The target we are estimating is the NE of a zero-sum MG, which is more challenging than standard statistical estimation problems in the sense that (i) NE is not unique in general and (ii) the error metric is a duality gap. It is challenging to use standard proof frameworks such as Fano's and Le Cam's methods to

derive a meaningful lower bound for this problem. To overcome this challenge, we construct a family of hard Markov game instances indexed by a binary parameter $\theta \in \{0,1\}^{\max\{A,B\}}$ and then put a prior distribution over this set and compute the posterior probability of failure to differentiate each entry of θ. These steps taken together carefully allow us to compute the desired minimax risk.

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

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

then one can construct a hard Markov game instance such that no algorithm whatsoever can reach a duality gap below ε. This, taken collectively (18), 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) propose 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

$$\tilde{O}\left(\frac{C^\star SABH^3}{\varepsilon^2}\right) \tag{19}$$

sample trajectories provided that $\varepsilon \le 1/H$. 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- 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 in two-player, zero-sum Markov games.
- 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 equivalence).
- The algorithm design herein is substantially simpler than Cui and Du (2022b): it neither requires sample splitting to decouple statistical dependency, nor relies on reference-advantage decomposition techniques to sharpen the horizon dependency.

When 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 nonstationary transition kernels, Cui and Du (2022a, algorithm 1) provably yields an ε-Nash policy pair using

$$\tilde{O}\left(\frac{C^\star S(A+B)H^4}{\varepsilon^2}\right) \tag{20}$$

sample trajectories, each containing H samples. This bound (20) 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, multiagent, general-sum MGs although the algorithm proposed therein becomes computationally intractable when going beyond two-player, zero-sum MGs.

4. Related Works

4.1. 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 Kumar et al. (2020), Kidambi et al. (2020), Yu et al. (2020; 2021a, b), Yin et al. (2021a, c), Rashidinejad et al. (2021), Jin et al. (2021b), Xie et al. (2021), Liu et al. (2020), Zhang et al. (2021c), Chang et al. (2021), Yin and Wang (2021), Uehara and Sun (2021), Munos (2003, 2007), Zanette et al. (2021), Yan et al. (2023), Li et al. (2022a, 2024b), Shi et al. (2022), Cui and Du (2022b), Zhong et al. (2022), Lu et al. (2022), Wang et al. (2022), and Xu and Liang (2022), which unveils 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), and Zhong et al. (2022), both of which focus on episodic, finite-horizon, zero-sum Markov games with two players. More specifically, Cui and Du (2022b) demonstrate that a unilateral concentrability condition is necessary for learning NEs in offline settings and propose a pessimistic value iteration with reference-advantage decomposition to enable sample efficiency. Zhong et al. (2022) propose a pessimistic minimax value iteration algorithm, which achieves appealing sample complexity in the presence of linear function representation and was recently improved by Xiong et al. (2022). In the concurrent work, Cui and Du (2022a) propose a different pessimistic algorithm that designs lower confidence bounds 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, Rashidinejad et al.

(2021), Yan et al. (2023), and Li et al. (2024b) study offline RL for infinite-horizon MDPs, and Jin et al. (2021b), Xie et al. (2021), Shi et al. (2022), and Li et al. (2024b) look at the finite-horizon episodic counterpart, all of which operate upon some single-policy concentrability assumptions. Among these works, Li et al. (2024b) and Yan et al. (2023) achieved minimax-optimal sample complexity $\tilde{O}\left(\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}\right)$ 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 (Xie et al. 2021; Yin et al. 2021b, c; Shi et al. 2022; Li et al. 2024b).

4.2. Multiagent RL and Markov Games

The concept of Markov games—also under the name of stochastic games—dates back to Shapley (1953), which has become a central framework to model competitive multiagent decision making. A large strand of prior works studies how to efficiently solve Markov games when perfect model description is available (Littman 1994, 2001; Hu and Wellman 2003; Hansen et al. 2013; Perolat et al. 2015; Daskalakis et al. 2020, 2023; Cen et al. 2021; Wei et al. 2021; Zhao et al. 2021; Chen et al. 2022; Mao and Başar 2022). 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, Wei et al. (2017), Xie et al. (2020), Bai et al. (2020), Bai and Jin (2020), Liu et al. (2021), Jin et al. (2021a), Song et al. (2021), Mao and Başar (2022), Daskalakis et al. (2023), Tian et al. (2021), and Chen et al. (2022) focus on the online explorative environments, whereas Zhang et al. (2020a) pays attention to the scenario that assumes sampling access to a generative model. Whereas the majority of these works exhibits a sample complexity that scales at least as $\tilde{O}(SAB)$ in order to learn an approximate NE, the recent work Bai et al. (2020) proposes 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. (2022b) further develops an algorithm that learns ε-Nash using $\tilde{O}\left(\frac{H^4S(A+B)}{\varepsilon^2}\right)$ samples, which attains the minimax lower bound for nonstationary, finite-horizon MGs. The setting of general-sum, multiplayer Markov games is much more challenging given that learning Nash equilibria is known to be PPAD-complete (Daskalakis et al. 2009, Daskalakis 2013). Shifting attention to more tractable solution concepts, Jin et al. (2021a), Daskalakis et al. (2023), Mao and Başar (2022), and Song et al. (2021) propose 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 ith player), thereby breaking the curse of multiagents. Additionally, there are also several works investigating the turn-based setting in which the two players take actions in turn; see Sidford et al. (2020), Cui and Yang (2021), Jia et al. (2019), and Jin et al. (2022). Moreover, another two works Zhang et al. (2021b) and Abe and Kaneko (2020) study offline sampling oracles under uniform coverage requirements (which are clearly more stringent than the unilateral concentrability assumption). The interested readers are also referred to Zhang et al. (2021a) and Yang and Wang (2020) for an overview of recent development.

4.3. 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 is extensively studied in the single-agent setting, including the online exploration setting (Azar et al. 2017, Zhang et al. 2023), the case with a generative model (Azar et al. 2013, Agarwal et al. 2020, Jin and Sidford 2021, Wang et al. 2021, Li et al. 2024a), the offline RL setting (Xie et al. 2021, Li et al. 2024b), 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., Azar et al. 2017, Agarwal et al. 2020, Zhang et al. 2023, Li et al. 2024b), sometimes even without incurring any burn-in cost (Cui and Yang 2021; Zhang et al. 2023; Li et al. 2024a, b). The method proposed in Cui and Du (2022b) also exhibits the flavor of a model-based algorithm although an additional variance reduction scheme is incorporated in order to optimize the horizon dependency.

5. Additional Notation

Let us collect a set of additional notations that are 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

$$\mathsf{Var}_{P_{s,a,b}}(V) = P_{s,a,b}(V \circ V) - (P_{s,a,b}V)^2, \tag{21}$$

where $P_{s,a,b}$ abbreviates $P(\cdot|s,a,b)$ as usual. 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 (Q-function) for all states (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 is often 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 denote by $\mu_{\mathsf{br}}(\nu)$. These allow one to define

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

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

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

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.

6. Proof of Theorem 1

Toward 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 $(\hat{\mu},\hat{\nu})$ returned by Algorithm 1 satisfies*

$$V^{\star}(\rho) - V^{\hat{\mu},\star}(\rho) \le c_0\sqrt{\frac{C^{\star}_{\mathsf{clipped}}S(A+B)}{(1-\gamma)^3N}\log\frac{N}{\delta}} + c_0\frac{C^{\star}_{\mathsf{clipped}}S(A+B)}{(1-\gamma)^2N}\log\frac{N}{\delta}, \quad (22a)$$

$$V^{\star,\hat{\nu}}(\rho) - V^{\star}(\rho) \le c_0\sqrt{\frac{C^{\star}_{\mathsf{clipped}}S(A+B)}{(1-\gamma)^3N}\log\frac{N}{\delta}} + c_0\frac{C^{\star}_{\mathsf{clipped}}S(A+B)}{(1-\gamma)^2N}\log\frac{N}{\delta} \quad (22b)$$

for some sufficiently large constant $c_0 > 0$. As an immediate consequence, the duality gap of $(\hat{\mu},\hat{\nu})$ obeys, with probability at least $1-\delta$, that

$$V^{\star,\hat{\nu}}(\rho) - V^{\hat{\mu},\star}(\rho) \le 2c_0\sqrt{\frac{C^{\star}_{\mathsf{clipped}}S(A+B)}{(1-\gamma)^3N}\log\frac{N}{\delta}} + 2c_0\frac{C^{\star}_{\mathsf{clipped}}S(A+B)}{(1-\gamma)^2N}\log\frac{N}{\delta}. \quad (23)$$

As can be straightforwardly verified, Theorem 1 is a direct consequence of Theorem 3 (by taking the right-hand side of (23) to be no larger than ε).

The remainder of this section is, thus, dedicated to establishing Theorem 3. Before proceeding, let us now take a moment to provide a brief road map of the proof.

1. We first show in Section 6.2 that the pessimistic Bellman operators $\hat{\mathcal{T}}^{-}_{\mathsf{pe}}$ and $\hat{\mathcal{T}}^{+}_{\mathsf{pe}}$ introduced in (11) are both monotone and γ-contractive and admit unique fixed points $Q^{-\star}_{\mathsf{pe},t}$ and $Q^{-\star}_{\mathsf{pe},t}$, respectively. These properties reveal that the pessimistic value iterations $\{Q^{-}_{\mathsf{pe},t}\}_{1\le t\le T}$ ($\{Q^{+}_{\mathsf{pe},t}\}_{1\le t\le T}$) in Algorithm 1 converge to $Q^{-\star}_{\mathsf{pe},t}$ ($Q^{-\star}_{\mathsf{pe},t}$) at a geometric rate, and therefore, it suffices to analyze the fixed points $Q^{-\star}_{\mathsf{pe},t}$ and $Q^{+\star}_{\mathsf{pe},t}$.

2. Next, we show Bernstein-style concentration bounds for random quantities such as $(\hat{P}_{s,a,b} - P_{s,a,b})V^{-\star}_{\mathsf{pe},t}$ and $(\hat{P}_{s,a,b} - P_{s,a,b})V^{+\star}_{\mathsf{pe},t}$ in Section 6.3, in which $V^{-\star}_{\mathsf{pe},t}$ and $V^{+\star}_{\mathsf{pe},t}$ are the value functions associated with $Q^{-\star}_{\mathsf{pe},t}$ and $Q^{+\star}_{\mathsf{pe},t}$. Because of the complicated statistical dependency between $\hat{P}_{s,a,b}$ and $V^{-\star}_{\mathsf{pe},t}$, we use a leave-one-out argument to establish this concentration result in Lemma 2.

3. Finally, based on the aforementioned results, we derive error bounds for $V^{\star}(\rho) - V^{\hat{\mu},\star}(\rho)$ and $V^{\star,\hat{\nu}}(\rho) - V^{\star}(\rho)$ in Section 6.4. Our analysis makes use of a self-bounding trick, which allows one to derive sharp estimation error bounds that turn out to be minimax-optimal.

6.1. Preliminary Facts

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

1. For any $Q_1, Q_2 : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \mathbb{R}$, we have
$$\|V_1 - V_2\|_\infty \le \|Q_1 - Q_2\|_\infty, \quad (24)$$
where V_1 (V_2) denotes the value function associated with Q_1 (Q_2); see (10) 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
$$|\mathsf{Var}_{P_{s,a,b}}(V_1) - \mathsf{Var}_{P_{s,a,b}}(V_2)| \le \frac{4}{1-\gamma}\|V_1 - V_2\|_\infty, \quad (25)$$
where $\mathsf{Var}_{P,s,a,b}(V)$ is defined in (21).

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. (12)) obey
$$|\beta(s,a,b;V_1) - \beta(s,a,b;V_2)| \le 2\|V_1 - V_2\|_\infty. \quad (26)$$

The proof of the preceding results can be found in Online Appendix EC.1.1.

6.2. Step 1: Key Properties of Pessimistic Bellman Operators

Recall the definition of the pessimistic Bellman operators $\hat{\mathcal{T}}^{-}_{\mathsf{pe}}$ and $\hat{\mathcal{T}}^{+}_{\mathsf{pe}}$ introduced in (11). The following lemma gathers a couple of key properties of these two operators.

Lemma 1. *The following properties hold true:*

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

$$\|\hat{\mathcal{T}}^{-}_{\mathsf{pe}}(Q_1) - \hat{\mathcal{T}}^{-}_{\mathsf{pe}}(Q_2)\|_\infty \le \gamma\|Q_1 - Q_2\|_\infty,$$
$$\|\hat{\mathcal{T}}^{+}_{\mathsf{pe}}(Q_1) - \hat{\mathcal{T}}^{+}_{\mathsf{pe}}(Q_2)\|_\infty \le \gamma\|Q_1 - Q_2\|_\infty$$

for any Q_1 and Q_2.

• *(Uniqueness of fixed points)* $\hat{\mathcal{T}}^{-}_{\mathsf{pe}}$ *(*$\hat{\mathcal{T}}^{+}_{\mathsf{pe}}$*) has a unique fixed point* $Q^{-\star}_{\mathsf{pe}}$ *(*$Q^{+\star}_{\mathsf{pe}}$*), which also satisfies* $0 \le Q^{\star-}_{\mathsf{pe}}(s,a,b) \le \frac{1}{1-\gamma}$ *(*$0 \le Q^{\star+}_{\mathsf{pe}}(s,a,b) \le \frac{1}{1-\gamma}$*) for any* $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$.

Proof. See Online Appendix EC.1.2. □

Next, we make note of several immediate consequences of Lemma 1. Here and throughout, $V^{-\star}_{\mathsf{pe}}$ and $V^{+\star}_{\mathsf{pe}}$ are defined to be the value functions (see (10)) associated with $Q^{-\star}_{\mathsf{pe}}$ and $Q^{+\star}_{\mathsf{pe}}$, respectively.

• First, the preceding lemma implies that

$$Q^{-}_{\mathsf{pe},t} \le Q^{-\star}_{\mathsf{pe}} \quad (\forall t \ge 0) \quad \text{hence} \quad Q^{-}_{\mathsf{pe}} \le Q^{-\star}_{\mathsf{pe}}. \tag{27}$$

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

$$Q^{-}_{\mathsf{pe},t+1} = \hat{\mathcal{T}}^{-}_{\mathsf{pe}}(Q^{-}_{\mathsf{pe},t}) \le \hat{\mathcal{T}}^{-}_{\mathsf{pe}}(Q^{-\star}_{\mathsf{pe}}) = Q^{-\star}_{\mathsf{pe}},$$

from which (27) follows.

• In addition, the γ-contraction property in Lemma 1 leads to

$$\|V^{-}_{\mathsf{pe}} - V^{-\star}_{\mathsf{pe}}\|_\infty \le \|Q^{-}_{\mathsf{pe}} - Q^{-\star}_{\mathsf{pe}}\|_\infty \le \frac{1}{N}, \tag{28}$$

and, to justify this, observe that

$$\begin{aligned}\|Q^{-}_{\mathsf{pe},t} - Q^{-\star}_{\mathsf{pe}}\|_\infty &= \|\hat{\mathcal{T}}^{-}_{\mathsf{pe}}(Q^{-}_{\mathsf{pe},t-1}) - \hat{\mathcal{T}}^{-}_{\mathsf{pe}}(Q^{-\star}_{\mathsf{pe}})\|_\infty \\ &\le \gamma\|Q^{-}_{\mathsf{pe},t-1} - Q^{-\star}_{\mathsf{pe}}\|_\infty \\ &\le \cdots \le \gamma^t\|Q^{-}_{\mathsf{pe},0} - Q^{-\star}_{\mathsf{pe}}\|_\infty \le \frac{\gamma^t}{1-\gamma},\end{aligned}$$

which, together with $T = \lceil \frac{\log(N/(1-\gamma))}{\log(1/\gamma)} \rceil$ and (24), gives

$$\begin{aligned}\|V^{-}_{\mathsf{pe}} - V^{-\star}_{\mathsf{pe}}\|_\infty &\le \|Q^{-}_{\mathsf{pe}} - Q^{-\star}_{\mathsf{pe}}\|_\infty = \|Q^{-}_{\mathsf{pe},T} - Q^{-\star}_{\mathsf{pe}}\|_\infty \\ &\le \frac{\gamma^T}{1-\gamma} \le \frac{1}{N}.\end{aligned}$$

• A similar argument also yields

$$Q^{+}_{\mathsf{pe}} \ge Q^{+\star}_{\mathsf{pe}}, \quad \|Q^{+}_{\mathsf{pe}} - Q^{+\star}_{\mathsf{pe}}\|_\infty \le 1/N, \quad \|V^{+}_{\mathsf{pe}} - V^{+\star}_{\mathsf{pe}}\|_\infty \le 1/N. \tag{29}$$

6.3. Step 2: Decoupling Statistical Dependency and Establishing Pessimism

To proceed, we rely on the following theorem to quantify the difference between $\hat{P}$ 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* $N(s,a,b) \ge 1$ *with probability exceeding* $1-\delta$,

$$|(\hat{P}_{s,a,b} - P_{s,a,b})\tilde{V}| \le \tilde{c}\sqrt{\frac{1}{N(s,a,b)}\mathsf{Var}_{\hat{P}_{s,a,b}}(\tilde{V})\log\frac{N}{\delta}} + \tilde{c}\frac{\log\frac{N}{\delta}}{(1-\gamma)N(s,a,b)} \tag{30}$$

for some sufficiently large constant $\tilde{c} > 0$, *and*

$$\mathsf{Var}_{\hat{P}_{s,a,b}}(\tilde{V}) \le 2\mathsf{Var}_{P_{s,a,b}}(\tilde{V}) + O\left(\frac{1}{(1-\gamma)^2 N(s,a,b)}\log\frac{N}{\delta}\right) \tag{31}$$

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

The proof is deferred to Online Appendix EC.1.3.

In words, the first result (30) delivers some Bernstein-type concentration bound, whereas the second result (31) 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 $\tilde{V}$ to be statistically independent from $\hat{P}_{s,a,b}$, which is particularly crucial when coping with the complicated statistical dependency of our iterative algorithm. The proof of Lemma 2 is established upon a leave-one-out analysis argument (see, e.g., Chen et al. 2019a, b; Agarwal et al. 2020; Ma et al. 2020; Chen et al. 2021; Li et al. 2024a, b) that helps decouple statistical dependency; see details in Online Appendix EC.1.3. Armed with Lemma 2, we can readily see that

$$|(\hat{P}_{s,a,b} - P_{s,a,b})\tilde{V}| + \frac{4}{N} \le \beta(s,a,b;\tilde{V}) \tag{32}$$

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

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

$$Q^{-}_{\mathsf{pe}} \le Q^{\hat{\mu},\star}, \quad Q^{+}_{\mathsf{pe}} \ge Q^{\star,\hat{\nu}}, \quad V^{-}_{\mathsf{pe}} \le V^{\hat{\mu},\star} \quad \text{and} \quad V^{+}_{\mathsf{pe}} \ge V^{\star,\hat{\nu}}.$$

The proof is provided in Online Appendix EC.1.4.

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

6.4. Step 3: Bounding $V^{\star}(\boldsymbol{\rho}) - V^{\hat{\mu},\star}(\boldsymbol{\rho})$ and $V^{\star,\hat{\nu}}(\boldsymbol{\rho}) - V^{\star}(\boldsymbol{\rho})$

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

$$\begin{aligned}V^{-}_{\mathsf{pe}}(s) &= \max_{\mu(s)\in\Delta(\mathcal{A})}\min_{\nu(s)\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim\mu(s),b\sim\nu(s)}[Q^{-}_{\mathsf{pe}}(s,\cdot,\cdot)] \\ &\ge \min_{\nu(s)\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim\mu^\star(s),b\sim\nu(s)}[Q^{-}_{\mathsf{pe}}(s,a,b)],\end{aligned} \tag{33}$$

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

$$\nu_0(s) = \arg\min_{\nu(s)\in\Delta(\mathcal{B})} \mathop{\mathbb{E}}_{a\sim\mu^\star(s), b\sim\nu(s)} [Q^-_{\mathsf{pe}}(s,a,b)] \tag{34}$$

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

$$\nu_0(s) = \mathbb{1}_{b_s} \qquad \text{with } b_s := \arg\max_{b\in\mathcal{B}} \langle \mu^\star(s), Q^-_{\mathsf{pe}}(s,\cdot,b)\rangle, \tag{35}$$

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 as follows, 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 $s, s' \in \mathcal{S}$,

$$r^{\mu^\star,\nu_0}(s) := \mathop{\mathbb{E}}_{a\sim\mu^\star(s), b\sim\nu_0(s)} [r(s,a,b)], \tag{36a}$$

$$\beta^{\mu^\star,\nu_0}(s) := \mathop{\mathbb{E}}_{a\sim\mu^\star(s), b\sim\nu_0(s)} [\beta(s,a,b;V^-_{\mathsf{pe}})], \tag{36b}$$

$$P^{\mu^\star,\nu_0}(s'|s) := \mathop{\mathbb{E}}_{a\sim\mu^\star(s), b\sim\nu_0(s)} [P(s'|s,a,b)]. \tag{36c}$$

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

$$V^-_{\mathsf{pe}} \geq r^{\mu^\star,\nu_0} + \gamma P^{\mu^\star,\nu_0} V^-_{\mathsf{pe}} - 2\beta^{\mu^\star,\nu_0}. \tag{37}$$

The proof is deferred to Online Appendix EC.1.5.

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

$$V^\star - V^{\hat{\mu},\star} = V^{\mu^\star,\star} - V^{\hat{\mu},\star} \leq V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}}, \tag{38}$$

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

$$V^{\mu^\star,\nu_0} = r^{\mu^\star,\nu_0} + \gamma P^{\mu^\star,\nu_0} V^{\mu^\star,\nu_0}. \tag{39}$$

Taking (37) and (39) collectively yields

$$V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}} \leq \gamma P^{\mu^\star,\nu_0}(V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}}) + 2\beta^{\mu^\star,\nu_0}, \tag{40}$$

thus resulting in a self-bounding type of relation. Applying (40) recursively, we arrive at that

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

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))

$$\begin{aligned}
d^{\mu^\star,\nu_0} &= (1-\gamma)\rho^\top\sum_{i=0}^{\infty}\gamma^i(P^{\mu^\star,\nu_0})^i \\
&= (1-\gamma)\rho^\top(I - \gamma P^{\mu^\star,\nu_0})^{-1},
\end{aligned} \tag{41}$$

we arrive at

$$\begin{aligned}
\rho^\top(V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}}) &\leq \{\lim_{n\to\infty}\ \gamma^n\rho^\top(P^{\mu^\star,\nu_0})^n(V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}})\} \\
&\quad + \frac{2}{1-\gamma}(d^{\mu^\star,\nu_0})^\top\beta^{\mu^\star,\nu_0} \\
&= \frac{2}{1-\gamma}(d^{\mu^\star,\nu_0})^\top\beta^{\mu^\star,\nu_0},
\end{aligned} \tag{42}$$

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

In order to further control (42), we resort to the following lemma for bounding $(d^{\mu^\star,\nu_0})^\top\beta^{\mu^\star,\nu_0}$, whose proof can be found in Online Appendix EC.1.6.

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

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

This is with probability exceeding $1-\delta$.

To finish, taking (38), (42) and Lemma 5 together gives

$$\begin{aligned}
V^\star(\rho) - V^{\hat{\mu},\star}(\rho) = \rho^\top(V^\star - V^{\hat{\mu},\star}) &\leq \rho^\top(V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}}) \\
&\leq \frac{2}{1-\gamma}(d^{\mu^\star,\nu_0})^\top\beta^{\mu^\star,\nu_0} \\
&\leq 2c_6\sqrt{\frac{C^\star_{\mathsf{clipped}}S(A+B)}{N(1-\gamma)^3}\log\frac{N}{\delta}} \\
&\quad + \frac{2c_6C^\star_{\mathsf{clipped}}S(A+B)}{(1-\gamma)^2N}\log\frac{N}{\delta}.
\end{aligned}$$

This completes the proof for Claim (22a). The proof for the other claim (22b) follows from an almost identical argument and is, hence, omitted.

6.5. Discussion: Instance-Dependent Statistical Bounds?

Thus far, we have presented the proof of Theorem 1 that concerns the minimax optimality of the model-based algorithm. Note that a recent line of work has attempted to move beyond minimax-optimal statistical guarantees and pursue more refined instance-optimal (or locally minimax) performance guarantees (Khamaru et al. 2021a, b, Mou et al. 2022). Here, we take a moment to discuss the challenges that need to be overcome in order to extend our analysis in an instance-optimal fashion.

- A crucial step that allows us to obtain error bounds that scale linearly with $A + B$ instead of the ones that scale linearly with AB in Cui and Du (2022b) is the introduction of an auxiliary policy ν_0 in (34). This

allows us to upper bound the error with

$$\rho^\top(V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}}) \le \frac{2}{1-\gamma}(d^{\mu^\star,\nu_0})^\top \beta^{\mu^\star,\nu_0};$$

see (42). Whereas this facilitates our analysis, the terms $\beta^{\mu^\star,\nu_0}$ (defined in (36)) and $d^{\mu^\star,\nu_0}$ (defined in (41)) both depend on the auxiliary policy ν_0. Because of the complicated dependency between ν_0 (which is determined by a random function V^-_{pe}) and the model parameters, it remains quite challenging to connect these error terms with instance-dependent quantities (i.e., model parameters) without losing optimality.

- Note that we might be able to resolve this issue in a coarse way, for example, by taking the supremum over all possible policy ν:

$$\rho^\top(V^{\mu^\star,\nu_0} - V^-_{\mathsf{pe}}) \le \frac{2}{1-\gamma}\sup_{\nu\in\Delta(\mathcal{B})}(d^{\mu^\star,\nu})^\top \beta^{\mu^\star,\nu}. \tag{43}$$

Let us assume for the moment that this could work (despite the potential suboptimality of this error bound) and see what this lead to. By checking the proof of Lemma 5, we can see that, in order to upper bound (43) in an instance-optimal manner, it is important to relate $\mathsf{Var}_{P_{s,a,b}}(V^-_{\mathsf{pe}})$ to model parameters for all $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$. Ideally, we can replace V^-_{pe} with the value function associated with the Nash equilibrium $V^\star$. However, based on our current analysis framework, we can only show that $V^\star(\rho)$ and $V^-_{\mathsf{pe}}(\rho)$ are close, which is insufficient to guarantee the closeness of $\mathsf{Var}_{P_{s,a,b}}(V^-_{\mathsf{pe}})$ and $\mathsf{Var}_{P_{s,a,b}}(V^\star)$ for all $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$.

- As we briefly mention in Remark 2, prior literature Khamaru et al. (2021a), which establishes instance optimality for the optimal value estimation problem in single-agent RL under a generative model, requires one of the following two conditions: the optimal policy is unique or a sample complexity bound that depends on an optimality gap

$$\Delta := \min_{\pi\in\mathbf{\Pi}\setminus\mathbf{\Pi}^\star}\|Q^\star - r - \gamma P^\pi Q^\star\|_\infty, \tag{44}$$

where $Q^\star$ is the optimal Q-function, $\mathbf{\Pi}$ is the set of deterministic policies, and $\mathbf{\Pi}^\star$ is the set of optimal (deterministic) policies. However, neither condition has a direct analog in two-player, zero-sum Markov games: for the first one, this is because the Nash equilibrium of a zero-sum Markov game is not unique in general; for the second one, this is because the Nash equilibrium policy pair is usually random, and it is not clear how to define a nonzero optimality gap such as (44).

In view of these challenges, our current analysis framework remains incapable of deriving instance-optimal performance guarantees. Accomplishing instance-optimal results for zero-sum Markov games might require substantially more refined analysis techniques, and we leave this important direction to future investigation.

7. Discussion

In the present paper, we propose 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 $\tilde{O}\left(\frac{S(A+B)C^\star}{(1-\gamma)^3\varepsilon^2}\right)$ samples, and this sample complexity bound is provably minimax optimal for the entire range of target accuracy level $\varepsilon\in\left(0,\frac{1}{1-\gamma}\right]$. Our theory improves 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 multiagent, general-sum Markov games with the goal of learning other solution concepts of equilibria such as 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, multiagent algorithms in the presence of function approximation.

Acknowledgments

Y. Chen thanks Shicong Cen for helpful discussions about Markov games.

References

Abe K, Kaneko Y (2020) Off-policy exploitability-evaluation in two-player zero-sum markov games. Preprint, submitted July 4, https://arxiv.org/abs/2007.02141.

Agarwal A, Kakade S, Yang LF (2020) Model-based reinforcement learning with a generative model is minimax optimal. Jacob A, Shivani A, eds. *Proc. 33rd Conf. Learn. Theory* (PMLR, New York), 67–83.

Azar MG, Munos R, Kappen HJ (2013) Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learn.* 91(3):325–349.

Azar MG, Osband I, Munos R (2017) Minimax regret bounds for reinforcement learning. Doina P, Yee Whye T, eds. *Proc. 34th Internat. Conf. Machine Learn.* (PMLR, New York), 263–272.

Bai Y, Jin C (2020) Provable self-play algorithms for competitive reinforcement learning. Daumé H III, Aarti S, eds. *Proc. 37th Internat. Conf. Machine Learn.* (PMLR, New York), 551–560.

Bai Y, Jin C, Yu T (2020) Near-optimal reinforcement learning with self-play. Larochelle H, Ranzato M, Hadsell R, Balcan MF,

Lin H, eds. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 2159–2170.

Baker B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, Mordatch I (2019) Emergent tool use from multi-agent autocurricula. *Internat. Conf. Learn. Representations* (Curran Associates, Inc., Red Hook, NY).

Berner C, Brockman G, Chan B, Cheung V, Debiak P, Dennison C, Farhi D, et al. (2019) Dota 2 with large scale deep reinforcement learning. Preprint, submitted December 13, https://arxiv.org/abs/1912.06680.

Bertsekas DP (2017) *Dynamic Programming and Optimal Control*, 4th ed. (Athena Scientific, Belmont, MA).

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

Cen S, Wei Y, Chi Y (2021) Fast policy extragradient methods for competitive games with entropy regularization. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. *Adv. Neural Inform. Processing Systems*, vol. 34 (Curran Associates, Inc., Red Hook, NY), 27952–27964.

Chang JD, Uehara M, Sreenivas D, Kidambi R, Sun W (2021) Mitigating covariate shift in imitation learning via offline data without great coverage. Preprint, submitted June 6, https://arxiv.org/abs/2106.03207.

Chen Z, Zhou D, Gu Q (2022) Almost optimal algorithms for two-player zero-sum linear mixture Markov games. Sanjoy D, Nika H, eds. *Proc. 33rd Internat. Conf. Algorithmic Learn. Theory* (PMLR, New York), 227–261.

Chen Y, Chi Y, Fan J, Ma C (2019a) Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Math. Programming* 176:5–37.

Chen Y, Chi Y, Fan J, Ma C (2021) Spectral methods for data science: A statistical perspective. *Foundations Trends Machine Learn.* 14(5): 566–806.

Chen Y, Fan J, Ma C, Yan Y (2019b) Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* 116(46):22931–22937.

Cui Q, Du SS (2022a) Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Adv. Neural Inform. Processing Systems*, vol. 35 (Curran Associates, Inc., Red Hook, NY), 11739–11751.

Cui Q, Du SS (2022b) When are offline two-player zero-sum Markov games solvable? Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Adv. Neural Inform. Processing Systems*, vol. 35 (Curran Associates, Inc., Red Hook, NY), 25779–25791.

Cui Q, Yang LF (2021) Minimax sample complexity for turn-based stochastic game. de Campos C, Maathuis MH, eds. *Uncertainty Artificial Intelligence* (PMLR, New York), 1496–1504.

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

Daskalakis C, Foster DJ, Golowich N (2020) Independent policy gradient methods for competitive reinforcement learning. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 5527–5540.

Daskalakis C, Goldberg PW, Papadimitriou CH (2009) The complexity of computing a Nash equilibrium. *SIAM J. Comput.* 39(1): 195–259.

Daskalakis C, Golowich N, Zhang K (2023) The complexity of Markov equilibrium in stochastic games. Gergely N, Lorenzo R, eds. *Proc. 36th Annual Conf. Learn. Theory* (PMLR, New York), 4180–4234.

Freund Y, Schapire RE (1999) Adaptive game playing using multiplicative weights. *Games Econom. Behav.* 29(1–2):79–103.

Hansen TD, Miltersen PB, Zwick U (2013) Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J. ACM* 60(1):1–16.

Hu J, Wellman MP (2003) Nash Q-learning for general-sum stochastic games. *J. Machine Learn. Res.* 4:1039–1069.

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

Jia Z, Yang LF, Wang M (2019) Feature-based q-learning for two-player stochastic games. Preprint, submitted June 2, https://arxiv.org/abs/1906.00423.

Jin Y, Sidford A (2021) Toward tight bounds on the sample complexity of average-reward MDPs. Marina M, Tong Z, eds. *Proc. 38th Internat. Conf. Machine Learn.* (PMLR, New York), 5055–5064.

Jin Y, Muthukumar V, Sidford A (2022) The complexity of infinite-horizon general-sum stochastic games. Preprint, submitted April 8, https://arxiv.org/abs/2204.04186.

Jin Y, Yang Z, Wang Z (2021b) Is pessimism provably efficient for offline RL? Marina M, Tong Z, eds. *Proc. 38th Internat. Conf. Machine Learn.* (PMLR, New York), 5084–5096.

Jin C, Liu Q, Wang Y, Yu T (2021a) V-learning—A simple, efficient, decentralized algorithm for multiagent RL. Preprint, submitted October 27, https://arxiv.org/abs/2110.14555.

Khamaru K, Xia E, Wainwright MJ, Jordan MI (2021a) Instance-optimality in optimal value estimation: Adaptivity via variance-reduced q-learning. Preprint, submitted June 28, https://arxiv.org/abs/2106.14352.

Khamaru K, Pananjady A, Ruan F, Wainwright MJ, Jordan MI (2021b) Is temporal difference learning optimal? An instance-dependent analysis. *SIAM J. Math. Data Sci.* 3(4):1013–1040.

Kidambi R, Rajeswaran A, Netrapalli P, Joachims T (2020) MOReL: Model-based offline reinforcement learning. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 21810–21823.

Kumar A, Zhou A, Tucker G, Levine S (2020) Conservative Q-learning for offline reinforcement learning. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 1179–1191.

Lagoudakis MG, Parr R (2002) Value function approximation in zero-sum Markov games. *Proc. 18th Conf. Uncertainty Artificial Intelligence* (Morgan Kaufmann Publishers Inc., Burlington, MA), 283–292.

Levine S, Kumar A, Tucker G, Fu J (2020) Offline reinforcement learning: Tutorial, review, and perspectives on open problems. Preprint, submitted May 4, https://arxiv.org/abs/2005.01643.

Li G, Ma C, Srebro N (2022a) Pessimism for offline linear contextual bandits using ℓ_p confidence sets. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Adv. Neural Inform. Processing Systems*, vol. 35 (Curran Associates, Inc., Red Hook, NY), 20974–20987.

Li G, Chi Y, Wei Y, Chen Y (2022b) Minimax-optimal multi-agent RL in zero-sum Markov games with a generative model. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 15353–15367.

Li G, Wei Y, Chi Y, Chen Y (2024a) Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Oper. Res.* 72(1):203–221.

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

Li G, Shi L, Chen Y, Gu Y, Chi Y (2021) Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. *Adv. Neural Inform. Processing Systems*, vol. 34 (Curran Associates, Inc., Red Hook, NY), 17762–17776.

Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *Machine Learn. Proc.* (Elsevier, Amsterdam), 157–163.

Littman ML (2001) Friend-or-foe Q-learning in general-sum games. *Proc. 18th Internat. Conf. Machine Learn.*, vol. 1 (Morgan Kaufmann Publishers Inc., Burlington, MA), 322–328.

Liu Y, Swaminathan A, Agarwal A, Brunskill E (2020) Provably good batch reinforcement learning without great exploration. Preprint, submitted July 16, https://arxiv.org/abs/2007.08202.

Liu Q, Yu T, Bai Y, Jin C (2021) A sharp analysis of model-based reinforcement learning with self-play. Marina M, Tong Z, eds. *Proc, 38th Internat. Conf. Machine Learn.* (PMLR, New York), 7001–7010.

Lu M, Min Y, Wang Z, Yang Z (2022) Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable Markov decision processes. Preprint, submitted May 26, https://arxiv.org/abs/2205.13589.

Ma C, Wang K, Chi Y, Chen Y (2020) Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations Comput. Math.* 20(3):451–632.

Mao W, Başar T (2022) Provably efficient reinforcement learning in decentralized general-sum Markov games. *Dynamic Games Appl.* 13:165–186.

Mou W, Khamaru K, Wainwright MJ, Bartlett PL, Jordan MI (2022) Optimal variance-reduced stochastic approximation in Banach spaces. Preprint, submitted January 21, https://arxiv.org/abs/2201.08518.

Munos R (2003) Error bounds for approximate policy iteration. *Proc. 20th Internat. Conf. Machine Learn.*, vol. 3 (AAAI Press, Washington, DC), 560–567.

Munos R (2007) Performance bounds in ℓ_p-norm for approximate value iteration. *SIAM J. Control Optim.* 46(2):541–561.

Nash J (1951) Non-cooperative games. *Ann. Math.* 54(2):286–295.

Perolat J, Scherrer B, Piot B, Pietquin O (2015) Approximate dynamic programming for two-player zero-sum Markov games. Francis B, David, eds. *Proc. 32nd Internat. Conf. Machine Learn.* (PMLR, New York), 1321–1329.

Raghavan T (1994) Zero-sum two-person games. Aumann R, Hart S, eds. *Handbook of Game Theory with Economic Applications*, vol. 2 (Elsevier, Amsterdam), 735–768.

Rakhlin S, Sridharan K (2013) Optimization, learning, and games with predictable sequences. Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems*, vol. 26 (Curran Associates, Inc., Red Hook, NY).

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

Roughgarden T (2016) *Twenty Lectures on Algorithmic Game Theory* (Cambridge University Press, Cambridge, MA).

Shalev-Shwartz S, Shammah S, Shashua A (2016) Safe, multi-agent, reinforcement learning for autonomous driving. Preprint, submitted October 11, https://arxiv.org/abs/1610.03295.

Shapley LS (1953) Stochastic games. *Proc. Natl. Acad. Sci. USA* 39(10):1095–1100.

Shi L, Li G, Wei Y, Chen Y, Chi Y (2022) Pessimistic Q-learning for offline reinforcement learning: Toward optimal sample complexity. Chaudhuri C, Jegelka S, Song L, Szepesvari C, Niu G, Sabato S, eds. *Proc. 39th Internat. Conf. Machine Learn.* (PMLR, New York), 19967–20025.

Sidford A, Wang M, Yang L, Ye Y (2020) Solving discounted stochastic two-player games with near-optimal time and sample complexity. Silvia C, Roberto C, eds. *Proc. 23rd Internat. Conf. Artificial Intelligence Statistics* (PMLR, New York), 2992–3002.

Song Z, Mei S, Bai Y (2021) When can we learn general-sum Markov games with a large number of players sample-efficiently? Preprint, submitted October 8, https://arxiv.org/abs/2110.04184.

Tian Y, Wang Y, Yu T, Sra S (2021) Online learning in unknown Markov games. Marina M, Tong Z, eds. *Internat. Conf. Machine Learn.* (PMLR, New York), 10279–10288.

Uehara M, Sun W (2021) Pessimistic model-based offline reinforcement learning under partial coverage. Preprint, submitted July 13, https://arxiv.org/abs/2107.06226.

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

Wang X, Cui Q, Du SS (2022) On gap-dependent bounds for offline reinforcement learning. Preprint, submitted June 1, https://arxiv.org/abs/2206.00177.

Wang B, Yan Y, Fan J (2021) Sample-efficient reinforcement learning for linearly-parameterized MDPs with a generative model. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. *Adv. Neural Inform. Processing Systems*, vol. 34 (Curran Associates, Inc., Red Hook, NY), 23009–23022.

Wei CY, Hong YT, Lu CJ (2017) Online reinforcement learning in stochastic games. Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 30 (Curran Associates, Inc., Red Hook, NY).

Wei CY, Lee CW, Zhang M, Luo H (2021) Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. Mikhail B, Samory K, eds. *Proc. 34th Conf. Learn. Theory* (PMLR, New York), 4259–4299.

Xie Q, Chen Y, Wang Z, Yang Z (2020) Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. Jacob A, Shivani A, eds. *Proc. 33rd Conf. Learn. Theory* (PMLR, New York), 3674–3682.

Xie T, Jiang N, Wang H, Xiong C, Bai Y (2021) Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. Preprint, submitted June 9, https://arxiv.org/abs/2106.04895.

Xiong W, Zhong H, Shi C, Shen C, Wang L, Zhang T (2022) Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent MDP and Markov game. Preprint, submitted May 31, https://arxiv.org/abs/2205.15512.

Xu T, Liang Y (2022) Provably efficient offline reinforcement learning with trajectory-wise reward. Preprint, submitted June 13, https://arxiv.org/abs/2206.06426.

Yan Y, Li G, Chen Y, Fan J (2023) The efficacy of pessimism in asynchronous Q-learning. *IEEE Trans. Inform. Theory* 69(11):7185–7219.

Yang Y, Wang J (2020) An overview of multi-agent reinforcement learning from game theoretical perspective. Preprint, submitted November 1, https://arxiv.org/abs/2011.00583.

Yin M, Wang YX (2021) Toward instance-optimal offline reinforcement learning with pessimism. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. *Adv. Neural Inform. Processing Systems*, vol. 34 (Curran Associates, Inc., Red Hook, NY).

Yin M, Bai Y, Wang YX (2021a) Near-optimal offline reinforcement learning via double variance reduction. Preprint, submitted February 2, https://arxiv.org/abs/2102.01748.

Yin M, Bai Y, Wang YX (2021b) Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. Arindam B, Kenji F, eds. *Proc. 24th Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 1567–1575.

Yin M, Duan Y, Wang M, Wang YX (2021c) Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *Internat. Conf. Learn. Representations*.

Yu T, Kumar A, Chebotar Y, Hausman K, Levine S, Finn C (2021a) Conservative data sharing for multi-task offline reinforcement learning. Preprint, submitted September 16, https://arxiv.org/abs/2109.08128.

Yu T, Kumar A, Rafailov R, Rajeswaran A, Levine S, Finn C (2021b) COMBO: Conservative offline model-based policy optimization. Preprint, submitted February 16, https://arxiv.org/abs/2102.08363.

Yu T, Thomas G, Yu L, Ermon S, Zou J, Levine S, Finn C, Ma T (2020) MOPO: Model-based offline policy optimization. Preprint, submitted May 27, https://arxiv.org/abs/2005.13239.

Zanette A, Wainwright MJ, Brunskill E (2021) Provable benefits of actor-critic methods for offline reinforcement learning. Preprint, submitted August 19, https://arxiv.org/abs/2108.08812.

Zhang K, Yang Z, Başar T (2021a) Multi-agent reinforcement learning: A selective overview of theories and algorithms. Vamvoudakis KG, Wan Y, Lewis FL, Cansever D, eds. *Handbook of Reinforcement Learning and Control* (Springer, Cham), 321–384.

Zhang Z, Zhou Y, Ji X (2020b) Almost optimal model-free reinforcement learning via reference-advantage decomposition. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 15198–15207.

Zhang Z, Chen Y, Lee JD, Du SS (2023) Settling the sample complexity of online reinforcement learning. Preprint, submitted July 25, https://arxiv.org/abs/2307.13586.

Zhang X, Chen Y, Zhu J, Sun W (2021c) Corruption-robust offline reinforcement learning. Preprint, submitted June 11, https://arxiv.org/abs/2106.06630.

Zhang K, Kakade S, Basar T, Yang L (2020a) Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Adv. Neural Inform. Processing Systems*, vol. 33 (Curran Associates, Inc., Red Hook, NY), 1166–1178.

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

Zhao Y, Tian Y, Lee JD, Du SS (2021) Provably efficient policy gradient methods for two-player zero-sum Markov games. Preprint, submitted February 17, https://arxiv.org/abs/2102.08903.

Zhong H, Xiong W, Tan J, Wang L, Zhang T, Wang Z, Yang Z (2022) Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. Preprint, submitted February 15, https://arxiv.org/abs/2202.07511.

Yuling Yan is a Norbert Wiener Postdoctoral Associate at the Massachusetts Institute of Technology. His research interests include statistics, optimization, and their applications in economics and social sciences. He has received the Institute of Mathematical Statistics Lawrence D. Brown Award, the Charlotte Elizabeth Procter Fellowship from Princeton University, the Norbert Wiener Postdoctoral Fellowship from MIT, and the International Consortium of Chinese Mathematicians Best Thesis Award.

Gen Li is an assistant professor of statistics at the Chinese University of Hong Kong. His research interests include reinforcement learning, high-dimensional statistics, machine learning, signal processing, and mathematical optimization. He has received the excellent graduate award and the excellent thesis award from Tsinghua University.

Yuxin Chen is an associate professor of statistics and data science at the University of Pennsylvania. His research interests include statistics, optimization, and machine learning. He has received the Alfred P. Sloan Research Fellowship, the Society for Industrial and Applied Mathematics Activity Group on Imaging Science Best Paper Prize, the International Consortium of Chinese Mathematicians Best Paper Award, and the Princeton Graduate Mentoring Award.

Jianqing Fan is the Frederick L. Moore Professor of Finance and Professor of Statistics at Princeton University. He was the past president of the Institute of Mathematical Statistics, and has received many awards and honors such as 2000 Committee of Presidents of Statistical Societies Presidents' Award, Guggenheim Fellow, Academician of Academia Sinica, and Royal Flemish Academy of Belgium. His research interests include statistics, data science, machine learning, and financial econometrics.