Journal of Machine Learning Research 24 (2023) 1-53 Submitted 10/20; Revised 7/23; Published 8/23

Model-Based Multi-Agent RL in Zero-Sum Markov Games with Near-Optimal Sample Complexity

Kaiqing Zhang kaiqing@umd.edu
University of Maryland, College Park
College Park, MD 20740, USA

Sham M. Kakade sham@seas.harvard.edu
Harvard University
Cambridge, MA 02138, USA

Tamer Başar basar1@illinois.edu
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Lin F. Yang linyang@ee.ucla.edu
University of California, Los Angeles
Los Angeles, CA 90095, USA

Editor: Joelle Pineau

Abstract

Model-based reinforcement learning (RL), which finds an optimal policy after establishing an empirical model, has long been recognized as one of the cornerstones of RL. It is especially suitable for multi-agent RL (MARL), as it naturally decouples the *learning* and the *planning* phases, and avoids the *non-stationarity* problem when all agents are improving their policies simultaneously. Though intuitive and widely-used, the sample complexity of model-based MARL algorithms has not been fully investigated. In this paper, we aim to address the fundamental question about its sample complexity. We study arguably the most basic MARL setting: two-player discounted zero-sum Markov games, given only access to a generative model. We show that model-based MARL achieves a sample complexity of $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-3}\epsilon^{-2})$ for finding the Nash equilibrium (NE) *value* up to some ϵ error, and the ϵ-NE *policies* with a smooth planning oracle, where γ is the discount factor, and $\mathcal{S},\mathcal{A},\mathcal{B}$ denote the state space, and the action spaces for the two agents. We further show that such a sample bound is *minimax-optimal* (up to logarithmic factors) if the algorithm is *reward-agnostic*, where the algorithm queries state transition samples without reward knowledge, by establishing a matching lower bound. This is in contrast to the usual *reward-aware* setting, where the sample complexity lower bound is $\widetilde{\Omega}(|\mathcal{S}|(|\mathcal{A}|+|\mathcal{B}|)(1-\gamma)^{-3}\epsilon^{-2})$, and this model-based approach is *near-optimal* with only a gap on the $|\mathcal{A}|,|\mathcal{B}|$ dependence. Our results not only illustrate the sample-efficiency of this basic model-based MARL approach, but also elaborate on the fundamental tradeoff between its power (easily handling the reward-agnostic case) and limitation (less adaptive and suboptimal in $|\mathcal{A}|,|\mathcal{B}|$), which particularly arises in the multi-agent context.

Keywords: Multi-Agent RL, Zero-Sum Markov Games, Near-Optimal Sample Complexity

1. Introduction

Recent years have witnessed numerous successes of reinforcement learning (RL) in many applications, e.g., playing strategy games (OpenAI, 2018; Vinyals et al., 2019), playing the game of Go (Silver et al., 2016, 2017), autonomous driving (Shalev-Shwartz et al., 2016), and security (Nguyen and Reddi, 2019; Zhang et al., 2019b). Most of these successful applications involve more than one decision-maker, giving birth to the surging interests and efforts in studying multi-agent RL (MARL) recently, especially on the theoretical side (Wei et al., 2017; Zhang et al., 2018a; Sidford et al., 2020; Zhang et al., 2019a; Xie et al., 2020; Shah et al., 2020; Bai and Jin, 2020; Bai et al., 2020). See also comprehensive surveys on MARL in Busoniu et al. (2008); Zhang et al. (2021a); Nguyen et al. (2020).

In general MARL, all agents affect both the state transition and the rewards of each other, while each agent may possess different, sometimes even totally conflicting objectives. Without knowledge of the model, the agents have to resort to data to either estimate the model, improve their own policy, and/or infer other agents' policies. One fundamental challenge in MARL is the emergence of *non-stationarity* during the learning process (Busoniu et al., 2008; Zhang et al., 2021a): when multiple agents improve their policies concurrently and directly using samples, the environment becomes non-stationary from each agent's perspective. This has posed great challenge to development of effective MARL algorithms based on single-agent ones, especially *model-free* ones, as the condition for guaranteeing convergence in the latter fails to hold in MARL. One tempting remedy for this non-stationarity issue is the simple while intuitive method — model-based[1] MARL: one first estimates an empirical model using data, and then finds the optimal, more specifically, equilibrium policies in this empirical model, via planning. Model-based MARL naturally decouples the *learning* and *planning* phases, and can be incorporated with *any* black-box planning algorithm that is efficient, e.g., value iteration (Shapley, 1953) and (generalized) policy iteration (Patek, 1997; Pérolat et al., 2015). More importantly, after estimating the model, this approach can potentially handle *more than one* MARL tasks with different reward functions but a common transition model, without re-sampling the data. Being able to handle this *reward-agnostic* case greatly expands the power of such a model-based approach.

Though intuitive and widely-used, rigorous theoretical justifications for these model-based MARL methods are relatively rare. In this work, our goal is to answer the following standing question: how good is the performance of this naïve "plug-in" method in terms of non-asymptotic sample complexity? To this end, we focus on arguably the most basic MARL setting since Littman (1994): two-player discounted zero-sum Markov games (MGs) with simultaneous-move agents, given only access to a generative model. This generative model allows agents to sample the MG, and query the next state from the transition process, given any state-action pair as input. The generative model setting has been a benchmark in RL when studying the sample efficiency of algorithms (Kearns and Singh, 1999; Kakade, 2003; Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2019a). Indeed, this model allows for the study of sample-based multi-agent planning over a long horizon, and helps

1. Note that we here follow the convention of *model-based* approach in the generative model setting (Azar et al., 2013; Agarwal et al., 2019a; Li et al., 2020), which separates these two stages explicitly. In general, model-based RL approaches do not have to separate the two stages, see e.g., Bayesian RL (Ghavamzadeh et al., 2015), and model-based RL in online exploration settings (Azar et al., 2017; Bai and Jin, 2020).

develop better understanding of the statistical properties of the algorithms, decoupled from the exploration complexity.

Motivated by recent minimax optimal complexity results for single-agent model-based RL (Agarwal et al., 2019a), we address the question above with a positive answer: the model-based MARL approach can achieve near-minimax optimal sample complexity — in terms of dependencies on the size of the state space, the horizon, and the desired accuracy — for finding both the Nash equilibrium (NE) value and the NE policies. We also provide a separation in the achievable sample complexity, unique to the multi-agent setting, where, with regards to the dependencies on the number of actions, the naïve model-based approach is sub-optimal. A detailed description is provided next.

Contributions. We establish the sample complexities of model-based MARL in zero-sum discounted Markov games, when a generative model is available. First, observing that the sampling process in this setting is agnostic to the reward function, we distinguish between two algorithmic frameworks: *reward-aware* and *reward-agnostic* cases, depending on whether the reward is revealed *before* or *after* the sampling. The model-based approach can inherently handle both cases, especially the latter case with multiple reward functions, without re-sampling the data. Second, by establishing lower bounds for both cases, we show that there is indeed a separation in sample complexity, which is unique in the multi-agent setting. Third, we show that up to some logarithmic factors, the model-based approach is indeed minimax optimal in all parameters in the more challenging reward-agnostic case, and has only a gap on the $|\mathcal{A}|, |\mathcal{B}|$ (both agents' action space size) dependence in the reward-aware case. This separation and the (near-)minimax results have not only justified the sample efficiency of this simple approach, but also highlighted both its power (easily handling multiple reward functions known in hindsight) and its limitation (less adaptive and can hardly achieve optimal complexity with reward knowledge), particularly arising in the multi-agent RL context. These results are first-of-their-kind in model-based MARL, and among the first (near-)minimax results in general MARL, to the best of our knowledge. We also believe that this separation may shed some light on the choice of model-free and model-based approaches in various MARL scenarios in practice, and provide new understandings for algorithm-design in other MARL settings, e.g., with no generative model, and going beyond two-player zero-sum MGs.

Related Work. Stemming from the formative work Littman (1994), MARL has been mostly studied under the framework of Markov games (Shapley, 1953). There has been no shortage of provably convergent MARL algorithms ever since then (Littman, 2001; Hu and Wellman, 2003; Greenwald et al., 2003). However, most of these early results are Q-learning-based (thus model-free) and asymptotic, with no sample complexity guarantees. To establish *non-asymptotic* results, Pérolat et al. (2015); Pérolat et al. (2016a,b); Fan et al. (2019); Zhang et al. (2018b) have studied the sample complexity of *batch* model-free MARL methods. There are also increasing interests in policy-based (thus also model-free) methods for solving special MGs with non-asymptotic convergence guarantees (Pérolat et al., 2018; Srinivasan et al., 2018; Zhang et al., 2019a). No result on the (near-)minimax optimality of these complexities has been established prior to the present work.

Specific to the two-player zero-sum setting, Jia et al. (2019) and Sidford et al. (2020) have considered *turn-based MGs*, a special case of the simultaneous-move MGs considered here, with a generative model. Specifically, Sidford et al. (2020) established near-optimal sample complexity of $\widetilde{\mathcal{O}}((1-\gamma)^{-3}\epsilon^{-2})$ for a variant of Q-learning for this setting. More recently, Bai and Jin (2020); Xie et al. (2020) have established both regret and sample complexity guarantees for episodic zero-sum MGs, without a generative model, with focus on efficient exploration. The work in Shah et al. (2020) also focused on the turn-based setting, and combined Monte-Carlo Tree Search and supervised learning to find the NE values. In contrast, model-based MARL theory has relatively limited literature. Brafman and Tennenholtz (2002) proposed the R-MAX algorithm for average-reward MGs, with polynomial sample complexity. Wei et al. (2017) developed a model-based upper confidence algorithm with polynomial sample complexities for the same setting. These methods differ from ours, as they are either specific model-free approaches, or not clear yet if they are (near-)minimax optimal in the corresponding setups. Concurrent to our work, Bai et al. (2020) developed *model-free* algorithms with near-optimal sample complexities in episodic settings without a generative model. The results are optimal in $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{B}|$ dependence, but not in the horizon H. Finally, we note that MARL in Markov games is not restricted to the competitive setting of two-player zero-sum, and the studies in (multi-player) cooperative/potential settings (Leonardos et al., 2021; Zhang et al., 2021b; Ding et al., 2022; Sayin et al., 2022) and general-sum settings (Hu and Wellman, 2003; Liu et al., 2021; Jin et al., 2021; Mao et al., 2022) also exist, and is not the focus of the present paper.

In the single-agent regime, there has been extensive literature on non-asymptotic efficiency of RL in MDPs; see Kearns and Singh (1999); Kakade (2003); Strehl et al. (2009); Jaksch et al. (2010); Azar et al. (2013); Osband and Van Roy (2014); Dann and Brunskill (2015); Azar et al. (2017); Wang (2017); Sidford et al. (2018); Jin et al. (2018); Li et al. (2020). Amongst them, we highlight the minimax optimal ones: Azar et al. (2013) and Azar et al. (2017) have provided minimax optimal results for sample complexity and regret in the settings with and without a generative model, respectively. Specifically, Azar et al. (2013) has shown that to achieve the ϵ-optimal *value* in Markov decision processes (MDPs), at least $\widetilde{\Omega}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\epsilon^{-2})$ samples are needed, for $\epsilon \in (0, 1]$. They also showed that to find an ϵ-optimal *policy*, the same minimax complexity order in $1-\gamma$ and ϵ can be attained, if $\epsilon \in (0, (1-\gamma)^{-1/2}|\mathcal{S}|^{-1/2}]$ and the total sample complexity is $\widetilde{\mathcal{O}}(|\mathcal{S}|^2|\mathcal{A}|)$, which is in fact *linear* in the model size. Later, Sidford et al. (2018) has proposed a Q-learning based approach to attain this lower bound and remove the extra dependence on $|\mathcal{S}|$, for $\epsilon \in (0, 1]$. More recently, Agarwal et al. (2019a) developed new techniques based on *absorbing MDPs*, to show that model-based RL also achieves the lower bound for finding an ϵ-optimal *policy*, with a larger ϵ range of $(0, (1-\gamma)^{-1/2}]$[2]. Finally, our separation of the reward-agnostic case is motivated by the recent novel framework of reward-free RL in Jin et al. (2020).

2. Preliminaries

Zero-Sum Markov Games. Consider a zero-sum MG[3] $\mathcal{G}$ characterized by $(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, \gamma)$, where $\mathcal{S}$ is the state space; $\mathcal{A}, \mathcal{B}$ are the action spaces of agents 1 and 2, respectively;

2. While preparing the present work, Li et al. (2020) has further improved the minimax optimal results in Agarwal et al. (2019a), in that they cover the entire range of sample sizes. We believe the improvement can also be incorporated in the MARL setting here, which is left as our future work.
3. We will hereafter refer to this model simply as a *MG*.

$P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ denotes the transition probability of states; $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ denotes the reward function[4] of agent 1 (thus $-r$ is the bounded reward function of agent 2); and $\gamma \in [0, 1)$ is the discount factor. The goal of agent 1 (agent 2) is to maximize (minimize) the long-term accumulative discounted reward. In MARL, the agents aim to achieve this goal using data samples collected from the model.

At each time t, agent 1 (agent 2) has a stationary (not necessarily deterministic) policy $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ ($\nu : \mathcal{S} \to \Delta(\mathcal{B})$), where $\Delta(\mathcal{X})$ denotes the space of all probability measures over $\mathcal{X}$, so that $a_t \sim \mu(\cdot \,|\, s_t)$ ($b_t \sim \nu(\cdot \,|\, s_t)$). The state makes a transition from s_t to s_{t+1} following the probability distribution $P(\cdot \,|\, s_t, a_t, b_t)$, given (a_t, b_t). As in the MDP model, one can define the *state-value function* under a pair of joint policies (μ, ν) as

$$V^{\mu,\nu}(s) := \mathbb{E}_{a_t \sim \mu(\cdot \,|\, s_t), b_t \sim \nu(\cdot \,|\, s_t)} \left[\sum_{t \geq 0} \gamma^t r(s_t, a_t, b_t) \,\middle|\, s_0 = s \right].$$

Note that $V^{\mu,\nu}(s) \in [0, 1/(1-\gamma)]$ for any $s \in \mathcal{S}$ as $r \in [0, 1]$, and the expectation is taken over the random trajectory produced by the joint policy (μ, ν). Also, the *state-action/Q-value function* under (μ, ν) is defined by

$$Q^{\mu,\nu}(s, a, b) := \mathbb{E}_{a_t \sim \mu(\cdot \,|\, s_t), b_t \sim \nu(\cdot \,|\, s_t)} \left[\sum_{t \geq 0} \gamma^t r(s_t, a_t, b_t) \,\middle|\, s_0 = s, a_0 = a, b_0 = b \right].$$

The solution concept considered is the (approximate) *Nash equilibrium*, as defined below.

Definition 1 ((ϵ-)Nash Equilibrium) *For a zero-sum MG $(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r, \gamma)$, a* Nash equilibrium policy *pair (μ^*, ν^*) satisfies the following pair of inequalities*[5] *for any $s \in \mathcal{S}$, $\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}$, and $\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}$*

$$V^{\mu,\nu^*}(s) \leq V^{\mu^*,\nu^*}(s) \leq V^{\mu^*,\nu}(s). \tag{1}$$

If (1) *holds with some $\epsilon > 0$ relaxation, i.e., for some policy (μ', ν'), such that*

$$V^{\mu,\nu'}(s) - \epsilon \leq V^{\mu',\nu'}(s) \leq V^{\mu',\nu}(s) + \epsilon, \tag{2}$$

then (μ', ν') is an ϵ-Nash equilibrium policy *pair.*

By Shapley (1953); Patek (1997), there exists a Nash equilibrium policy pair $(\mu^*, \nu^*) \in \Delta(\mathcal{A})^{|\mathcal{S}|} \times \Delta(\mathcal{B})^{|\mathcal{S}|}$ for two-player discounted zero-sum MGs. The state-value $V^* := V^{\mu^*,\nu^*}$ is referred to as the *value of the game*. The corresponding Q-value function is denoted by Q^*. The objective of the two agents is to find the NE policy of the MG, namely, to solve the saddle-point problem

$$\max_{\mu} \min_{\nu} \quad V^{\mu,\nu}(s), \tag{3}$$

for every $s \in \mathcal{S}$, where the order of max and min can be interchanged (Von Neumann et al., 1953; Shapley, 1953). For notational convenience, for any policy (μ, ν), we define

$$V^{\mu,*} = \min_{\nu} V^{\mu,\nu}, \qquad\qquad V^{*,\nu} = \max_{\mu} V^{\mu,\nu}, \tag{4}$$

and denote the corresponding optimizers by $\nu(\mu)$ and $\mu(\nu)$, respectively. We refer to these values and optimizers as the *best-response* values and policies, given μ and ν, respectively.

4. Our results can be generalized to other ranges of reward function by a standard reduction, see e.g., Sidford et al. (2018), and randomized reward functions.
5. In game theory, this pair is commonly referred to as *saddle-point inequalities*.

Reward-Aware v.s. Reward-Agnostic. We first differentiate between two algorithmic mechanisms in the generative-model setting. In the reward-aware case, the reward function is either known to the agents (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2019a; Sidford et al., 2020; Li et al., 2020), or can at least be estimated from data. The reward knowledge can thus be used to potentially guide the sampling process, making the algorithm *adaptive*. In the reward-agnostic case, reward knowledge is not used to guide sampling, and is possibly only revealed after the sampling. This especially fits in the scenario when there is more than one reward function of interest, or when the reward function is engineered iteratively, since it can now handle a class of reward functions that are not pre-specified, without re-sampling the data for each of them. Existing works in single-agent settings have no such a separation (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2019b; Li et al., 2020), as the sample complexity of estimating the reward function is typically of lower order, and the reward function is thus assumed to be known. In particular, the model-based approaches in Azar et al. (2013); Agarwal et al. (2019b); Li et al. (2020) are reward-agnostic, while the model-free approaches in Sidford et al. (2018, 2020) are reward-aware. Interestingly, in two-agent Markov games, whether the reward is known beforehand or not may lead to different sample complexity lower-bounds, as we will see in §3.1. We thus point out this separation here for clarity.

Remark 2 (Reward-Agnostic & Reward-Free) *The reward-agnostic case we advocate here is closely related to the recent novel algorithmic framework of* reward-free *RL (Jin et al., 2020), where there are also two phases,* exploration *and* planning*, while trajectories are only collected in the exploration phase, without any reward knowledge, and various reward functions are fed to the algorithm for evaluation in the planning phase. One key difference is that the reward-free setting aims to be effective for* all *reward function in the planning phase* simultaneously*, while the reward-agnostic setting only focuses on handling the underlying single-reward (or a few, e.g., polynomial number of, reward functions) that is not pre-specified. Being less general than the pure reward-free setting, the sample complexity bounds are thus possibly better, as we will show in §3.*

Model-Based Approach with Generative Model. As a standard setting, suppose that we have access to a *generative model/sampler*, which can provide us with samples $s' \sim P(\cdot \mid s, a, b)$ for any (s, a, b). The model-based MARL algorithm simply calls the sampler N times at each state-joint-action pair (s, a, b), and constructs an empirical estimate of the transition model P, denoted by $\widehat{P}$, following

$$\widehat{P}(s' \mid s, a, b) = \frac{\text{count}(s', s, a, b)}{N}.$$

Here $\text{count}(s', s, a, b)$ is the number of times the state-action pair (s, a, b) forces a transition to state s'. Note that the reward function is not estimated, in either the reward-aware or reward-agnostic cases, as for the former, the sample complexity of estimating r is only a lower order term, and r is thus typically assumed to be known (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2019a; Li et al., 2020); while for the latter, no reward information is even available in the sampling processes. This model-based approach via estimating P inherently handles both cases. Such a model-estimation can be implemented by both agents independently.

Planning Oracle. The reward function, together with the empirical transition model $\widehat{P}$ and the components $(\mathcal{S}, \mathcal{A}, \mathcal{B}, \gamma)$ in the true model $\mathcal{G}$, constitutes an empirical game model $\widehat{\mathcal{G}}$. As in Azar et al. (2013); Agarwal et al. (2019a); Jin et al. (2020); Li et al. (2020) for single-agent RL, we assume that an efficient planning oracle is available, which takes $\widehat{\mathcal{G}}$ as input, and outputs a policy pair $(\widehat{\mu}, \widehat{\nu})$. This oracle decouples the statistical and computational aspects of the empirical model $\widehat{\mathcal{G}}$. The output policy pair, referred to as being *near-equilibrium*, is assumed to satisfy certain ϵ_{opt}-order of equilibrium, in terms of value functions, and we evaluate the performance of $(\widehat{\mu}, \widehat{\nu})$ on the original MG $\mathcal{G}$. Common planning algorithms include value iteration (Shapley, 1953) and (generalized) policy iteration (Patek, 1997; Pérolat et al., 2015), which are efficient in finding the (ϵ-)NE of $\widehat{\mathcal{G}}$. In addition, it is not hard to have an oracle that is smooth in generating policies, i.e., the change of the approximate NE policies can be bounded by the changes of the NE value. See our Definition 7 later for a formal statement. Finally, we note that our definition of *model-based* approach in the generative-model-setting follows from that in (Azar et al., 2013; Agarwal et al., 2019a; Li et al., 2020), which separates these two stages explicitly. In general, model-based RL approaches do not have to separate the two stages, see e.g., Bayesian RL (Ghavamzadeh et al., 2015), and model-based RL in online exploration settings (Azar et al., 2017; Bai and Jin, 2020).

3. Main Results

We now introduce the main results of this paper. For notational convenience, we use $\widehat{V}^{\mu,\nu}$, $\widehat{V}^{\mu,*}$, $\widehat{V}^{*,\nu}$, and $\widehat{V}^{*}$ to denote the value under (μ, ν), the best-response value under μ and ν, and the NE value, under the empirical game model $\widehat{\mathcal{G}}$, respectively. A similar convention is also used for Q-functions.

3.1 Lower Bounds

We first establish lower bounds on both approximating the NE value function and learning the ϵ-NE policy pair, in both reward-aware and reward-agnostic cases.

Lemma 3 (Lower Bound for Reward-Aware Case) *Let $\mathcal{G}$ be an unknown zero-sum MG, and the agents learn in a reward-aware case, i.e., the reward knowledge is available during sampling. Then, there exist $\epsilon_0, \delta_0 > 0$, such that for all $\epsilon \in (0, \epsilon_0)$, $\delta \in (0, \delta_0)$, the sample complexity of learning an ϵ-NE policy pair, or an ϵ-approximate NE value, i.e., finding a $\widehat{Q}$ such that $\|\widehat{Q} - Q^*\|_\infty \leq \epsilon$ for $\mathcal{G}$, with a generative model with probability at least $1 - \delta$, is $\widetilde{\Omega}\big(|\mathcal{S}|(|\mathcal{A}| + |\mathcal{B}|)(1-\gamma)^{-3}\epsilon^{-2}\log(1/\delta)\big)$.*

The proof of Lemma 3, via a straightforward adaptation from the lower bounds for MDPs (Azar et al., 2013; Feng et al., 2019), is provided in §A.1. In particular, one can design a two-player zero-sum Markov game such that one of the players is dummy – she has no control on the reward nor the transition dynamics. Then, the existing lower bound in Azar et al. (2013); Feng et al. (2019) for MDPs leads to the desired result. Note that as in Azar et al. (2013); Sidford et al. (2018, 2020); Agarwal et al. (2019a); Li et al. (2020), the reward function is *known* in this case. As we will show momentarily, our sample complexity is tight in $1-\gamma$ and $|\mathcal{S}|$, while has a gap in $|\mathcal{A}|, |\mathcal{B}|$ dependence ($\widetilde{\mathcal{O}}(|\mathcal{A}||\mathcal{B}|)$ versus $\widetilde{\Omega}(|\mathcal{A}| + |\mathcal{B}|)$). In

§A.1, we discuss that the $\widetilde{\Omega}(|\mathcal{A}|+|\mathcal{B}|)$ lower bound may not be improved in this reward-aware case, and might be attainable by *model-free* algorithms instead (as $\widetilde{\mathcal{O}}(|\mathcal{A}||\mathcal{B}|)$ is inherent in model-based approaches due to estimating P). Interestingly, in the concurrent work Bai et al. (2020), under a different MARL setting, such an $\widetilde{\Omega}(|\mathcal{A}|+|\mathcal{B}|)$ complexity is indeed shown to be attainable by a model-free algorithm with *online* updates.

On the other hand, note that our model-based approach can inherently also handle the more challenging reward-agnostic case. Indeed, estimating the transition model P seems a bit of an overkill for the reward-aware case, in terms of sample complexity. A natural two-part question then becomes: what is the sample complexity lower bound in this more challenging reward-agnostic case, and can the model-based approach attain it? We formally answer the first part of the question in the following theorem, whose proof is deferred to §A.2, and answer the second part in §3.2 and §3.3.

Theorem 4 (Lower Bound for Reward-Agnostic Case) *Let $\mathcal{G}$ be an unknown zero-sum MG, and the agents learn in a reward-agnostic case, i.e., they first call the generative model for sampling, without reward knowledge, and then are fed with the reward function r in $\mathcal{G}$, for finding either an ϵ-NE policy pair, or an ϵ-approximate NE value for $\mathcal{G}$. Then, there exist $\epsilon_0, \delta_0 > 0$, such that for all $\epsilon \in (0, \epsilon_0)$, $\delta \in (0, \delta_0)$, the sample complexity of achieving either goal with probability at least $1-\delta$, is*

$$\widetilde{\Omega}\bigg(\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|}{(1-\gamma)^3\epsilon^2}\log\Big(\frac{1}{\delta}\Big)\bigg).$$

Compared to Lemma 3, the dependence on $|\mathcal{A}|, |\mathcal{B}|$ is increased from $\widetilde{\Omega}(|\mathcal{A}|+|\mathcal{B}|)$ to $\widetilde{\Omega}(|\mathcal{A}||\mathcal{B}|)$. Several remarks are now in order. First, this suggests that without guidance from the reward, the reward-agnostic case can be more challenging to tackle. The intuition is that, when the reward is only given in hindsight, which might be chosen adversarially, costs the algorithm to at least sample at all $|\mathcal{A}||\mathcal{B}|$ elements in the Q-value $Q(s,\cdot,\cdot)$ at each state s often enough. Second, when reduced to the single-agent setting (e.g., with $|\mathcal{B}| = 1$), such a separation disappears, showing its unique emergence in the multi-agent setting, and explaining why these two cases were not differentiated explicitly in the single-agent literature. Third, this lower bound is also related to the reward-free setting (Jin et al., 2020) with a single unknown reward (not infinitely many as in Jin et al. (2020)).

The basic intuition regarding the separation between the lower bounds in reward-aware and reward-agnostic cases, when compared to the single-agent setting (where there is no such a separation), is the *insensitivity* of Nash equilibrium (NE) to the changes in payoff matrices in two-player zero-sum games (Jansen, 1981). In particular, NE in general depends on the *joint* behavior and preferences of both agents. Specifically, to construct the lower bound (even in the single-agent case, see e.g., Azar et al. (2013); Feng et al. (2019)), we needed to carefully *perturb* the Q-value function at each state-action pair of some *null hypothesis* instance, so that the solution (which is the *maximum* in the single-agent case, and *Nash equilibrium* in the multi-agent case) is also changed in the perturbed alternative hypothesis cases. Hence, for each alternative hypothesis case, we need to change the NE by only changing $\mathcal{O}(1)$ elements in the payoff matrix, i.e., the Q-value table. In the reward-aware setting, since the reward is known (or can be estimated accurately with negligible sample complexity), we can *only perturb* the *transition matrix* to perturb the Q-value table, which share the same size (i.e., the degree-of-freedom). Due to the insensitivity, we can hardly construct $\Theta(|\mathcal{A}||\mathcal{B}|)$ different hard cases (as needed to construct a $\Omega(|\mathcal{A}||\mathcal{B}|)$ lower bound) while by only perturbing $\mathcal{O}(1)$ elements in the transition matrix in each case. Note that such a perturbation can be effective in the single-agent MDP setting, as by only perturbing one element in the transition matrix, the maximum of the Q-value can be changed, see e.g., Azar et al. (2013); Feng et al. (2019).

In contrast, in the reward-agnostic setting, the reward information is given *after* the sampling phase and the estimation of the model. This way, more freedom is allowed to construct $\Theta(|\mathcal{A}||\mathcal{B}|)$ different hard cases, by *adversarially choosing* the reward function afterwards. In particular, the Q-value will be affected by both the transition matrix and the reward, and with polynomial number of reward functions, we were able to construct Q-value tables in $\Theta(|\mathcal{A}||\mathcal{B}|)$ different hard cases. Note that taking a union bound over the polynomial number of reward functions do not affect the total sample complexity, as it is still dominated by the sample complexity of estimating the transition matrix. In other words, the freedom of constructing and perturbing the reward functions adversarially *afterwards* forces the algorithm to estimate *all* the elements in the transition matrix well, in order to handle the reward-agnostic setting. This has been inherently done by our model-based approach. We defer more details about the lower bounds comparison in Appendix A.

3.2 Near-Optimality in Finding ϵ-Approximate NE Value

We now establish the near-minimax optimal sample complexities of model-based MARL. Note that theses results apply to both reward-aware and reward-agnostic cases, as the implementation of our model-based approach does not rely on the reward function. We start by showing the sample complexity to achieve an ϵ-approximate NE value.

Theorem 5 (Finding ϵ-Approximate NE Value) *Suppose that the policy pair $(\widehat{\mu}, \widehat{\nu})$ is obtained from the **Planning Oracle** using the empirical model $\widehat{\mathcal{G}}$, which satisfies*

$$\|\widehat{V}^{\widehat{\mu},\widehat{\nu}} - \widehat{V}^*\|_\infty \leq \epsilon_{opt}.$$

Then, for any $\delta \in (0,1]$ and $\epsilon \in (0, 1/(1-\gamma)^{1/2}]$, if

$$N \geq \frac{c\gamma \log\left[c|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-2}\delta^{-1}\right]}{(1-\gamma)^3\epsilon^2}$$

for some absolute constant c, it holds that with probability at least $1-\delta$,

$$\left\|Q^{\widehat{\mu},\widehat{\nu}} - Q^*\right\|_\infty \leq \frac{2\epsilon}{3} + \frac{5\gamma\epsilon_{opt}}{1-\gamma}, \qquad \left\|\widehat{Q}^{\widehat{\mu},\widehat{\nu}} - Q^*\right\|_\infty \leq \epsilon + \frac{9\gamma\epsilon_{opt}}{1-\gamma}.$$

Theorem 5 shows that if the planning error ϵ_{opt} is made small, e.g., with the order of $\mathcal{O}((1-\gamma)\epsilon)$, then the Nash equilibrium Q-value can be estimated with a sample complexity of $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-3}\epsilon^{-2})$, as N queries are made for each (s,a,b) pair. This planning error can be achieved by performing any efficient black-box optimization technique over the empirical model $\widehat{\mathcal{G}}$. Examples of such oracles include value iteration (Shapley, 1953) and (generalized) policy iteration (Patek, 1997; Pérolat et al., 2015). Moreover, note that, in contrast to the single-agent setting, where only a max operator is used, a min max (or max min) operator is used in these algorithms, which involves solving a *matrix game* at each state. This can be solved as a linear program (Osborne and Rubinstein, 1994), with at best polynomial runtime complexity (Grötschel et al., 1981; Karmarkar, 1984). This in total leads to an efficient polynomial runtime complexity algorithm.

As per Lemma 3, our $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-3}\epsilon^{-2})$ complexity is near-minimax optimal for the reward-aware case, in that it is tight in the dependence of $1-\gamma$ and $|\mathcal{S}|$, and sublinear in the model-size (which is $|\mathcal{S}|^2|\mathcal{A}||\mathcal{B}|$). However, there is a gap on the $|\mathcal{A}|$, $|\mathcal{B}|$ dependence ($\widetilde{\mathcal{O}}(|\mathcal{A}||\mathcal{B}|)$ versus $\widetilde{\mathcal{O}}(|\mathcal{A}|+|\mathcal{B}|)$). Unfortunately, without further assumption on the MG, e.g., being turn-based, the model-based algorithm can hardly avoid the $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|)$ dependence, as it is required to estimate each $\widehat{P}(\cdot \mid s,a,b)$ accurately to perform the planning. It is only minimax-optimal if the action-space size of one agent dominates the other's (e.g., $|\mathcal{A}| \gg |\mathcal{B}|$).

In the reward-agnostic case, as per Theorem 4, $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-3}\epsilon^{-2})$ is indeed minimax-optimal, and is tight in all $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{B}|$ and $1-\gamma$ dependence. More significantly, in this case, more than one reward functions can be handled simultaneously, as long as the transition model is estimated accurately enough. Specifically, with M reward functions, by letting $\delta = \delta/M$ in Theorem 5 and using union bounds, the sample complexity of finding ϵ-approximate NE value corresponding to all M reward functions becomes $\widetilde{\mathcal{O}}(\log(M)|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-3}\epsilon^{-2})$, which, with M being polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, $|\mathcal{B}|$, is of the same order as that in Theorem 5.

However, this (near-)optimal result does not necessarily lead to near-optimal sample complexity for obtaining the ϵ-NE *policies*. We first use a direct translation to obtain such an ϵ-NE policy pair based on Theorem 5, for *any* **Planning Oracle**.

Corollary 6 (Finding ϵ-NE Policy) *Let $(\widehat{\mu},\widehat{\nu})$ and N satisfy the conditions in Theorem 5. Let*

$$\widetilde{\epsilon} := \frac{2}{1-\gamma}\cdot\Big(\epsilon + \frac{9\gamma\epsilon_{opt}}{1-\gamma}\Big),$$

and $(\widetilde{\mu},\widetilde{\nu})$ be the one-step Nash equilibrium of $\widehat{Q}^{\widehat{\mu},\widehat{\nu}}$, namely, for any $s \in \mathcal{S}$

$$\big(\widetilde{\mu}(\cdot \mid s), \widetilde{\nu}(\cdot \mid s)\big) \in \mathop{\mathrm{argmax}}_{u\in\Delta(\mathcal{A})} \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim u, b\sim\vartheta}\big[\widehat{Q}^{\widehat{\mu},\widehat{\nu}}(s,a,b)\big].$$

Then, with probability at least $1-\delta$,

$$V^{*,\widetilde{\nu}} - 2\widetilde{\epsilon} \leq V^{\widetilde{\mu},\widetilde{\nu}} \leq V^{\widetilde{\mu},*} + 2\widetilde{\epsilon}, \tag{5}$$

namely, $(\widetilde{\mu},\widetilde{\nu})$ constitutes a $2\widetilde{\epsilon}$-Nash equilibrium policy pair.

Corollary 6 is equivalently to saying that the sample complexity of achieving an ϵ-NE policy pair is $\widetilde{\mathcal{O}}((1-\gamma)^{-5}\epsilon^{-2})$. This is worse than the model-based single-agent setting (Agarwal et al., 2019a), and also worse than both the model-free single-agent (Sidford et al., 2018) and turn-based two-agent (Sidford et al., 2020) settings, where $\widetilde{\mathcal{O}}((1-\gamma)^{-3}\epsilon^{-2})$ can be achieved for learning the optimal policy. This also has a gap from the lower bound in both Lemma 3 and Theorem 4. Note that the above sample complexity still matches that of the Empirical QVI in Azar et al. (2013) if $\epsilon \in (0,1]$ for single-agent RL, but with a larger choice of ϵ of $(0,(1-\gamma)^{-1/2}]$. As the Markov game setting is more challenging than MDPs, it is not clear yet if the lower bounds in Lemma 3 and Theorem 4 in finding ϵ-NE policies can be achieved, using a general **Planning Oracle**. In contrast, we show next that a stable **Planning Oracle** can indeed (almost) match the lower bounds.

3.3 Near-Optimality in Finding ϵ-NE Policy

Admittedly, Corollary 6 does not fully exploit the *model-based* approach, since it finds the NE policy according to the Q-value estimate $\widehat{Q}^{\widehat{\mu},\widehat{\nu}}$, instead of using the output policy pair $(\widehat{\mu},\widehat{\nu})$ directly. This loses a factor of $1-\gamma$. To improve the sample complexity of obtaining the NE policies, we first introduce the following definition of a smooth **Planning Oracle**.

Definition 7 (Smooth Planning Oracle) *A smooth **Planning Oracle** generates policies that are smooth with respect to the NE Q-values of the empirical model. Specifically, for two empirical models $\widehat{\mathcal{G}}_1$ and $\widehat{\mathcal{G}}_2$, the generated near-equilibrium policy pair $(\widehat{\mu}_1,\widehat{\nu}_1)$ and $(\widehat{\mu}_2,\widehat{\nu}_2)$ satisfy that for each $s\in\mathcal{S}$, $\|\widehat{\mu}_1(\cdot\,|\,s)-\widehat{\mu}_2(\cdot\,|\,s)\|_{TV}\leq C\cdot\|\widehat{Q}^*_1-\widehat{Q}^*_2\|_\infty$ and $\|\widehat{\nu}_1(\cdot\,|\,s)-\widehat{\nu}_2(\cdot\,|\,s)\|_{TV}\leq C\cdot\|\widehat{Q}^*_1-\widehat{Q}^*_2\|_\infty$ for some constant[6] $C>0$, where $\widehat{Q}^*_i$ is the NE Q-value of $\widehat{\mathcal{G}}_i$ for $i=1,2$, and $\|\cdot\|_{TV}$ is the total variation distance.*

Such a smooth **Planning Oracle** can be readily obtained in several ways. For example, one simple (but possibly computationally expensive) approach is to output the average over the entire policy space, using a *softmax* randomization over best-response values induced by $\widehat{Q}^*$. Specifically, for agent 1, the output $\widehat{\mu}$ is given by

$$\widehat{\mu}(\cdot\,|\,s)=\int_{\Delta(\mathcal{A})}\frac{\exp\big(\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big]/\tau\big)}{\int_{\Delta(\mathcal{A})}\exp\big(\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u',b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big]/\tau\big)du'}\cdot u du,$$

where $\tau>0$ is some temperature constant. The output of $\widehat{\nu}$ is analogous. With a small enough τ, $\widehat{\mu}$ approximates the exact solution to $\operatorname{argmax}_{u\in\Delta(\mathcal{A})}\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}[\widehat{Q}^*(s,a,b)]$, the NE policy given $\widehat{Q}^*$. Moreover, notice that $\widehat{\mu}$ satisfies the smoothness condition in Definition 7. This is because for each $u\in\Delta(\mathcal{A})$ in the integral: i) the softmax function is Lipschitz continuous with respect to $\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big]$ with Lipschitz constant $1/\tau$ (Gao and Pavel, 2017); ii) the best-response value $\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big]$ is smooth with respect to $\widehat{Q}^*$. Thus, such an oracle is an instance of a smooth **Planning Oracle**.

Another more tractable way to obtain $(\widehat{\mu},\widehat{\nu})$ is by directly solving a *regularized* matrix game induced by $\widehat{Q}^*$. Specifically, one solves

$$\big(\widehat{\mu}(\cdot\,|\,s),\widehat{\nu}(\cdot\,|\,s)\big)=\operatorname*{argmax}_{u\in\Delta(\mathcal{A})}\min_{\vartheta\in\Delta(\mathcal{B})}\quad\mathbb{E}_{a\sim u,b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big]-\tau_1\Omega_1(u)+\tau_2\Omega_2(\vartheta),\tag{6}$$

for each $s\in\mathcal{S}$, where Ω_i is the regularizer for agent i's policy, usually a strongly convex function, $\tau_i>0$ are the temperature parameters. This strongly-convex-strongly-concave

6. We allow C to depend polynomially on $|\mathcal{A}|$, $|\mathcal{B}|$, which, as we will show later, does not affect the sample complexity as it appears as $\log C$.

saddle point problem admits a unique solution, and can be solved efficiently (Facchinei and Pang, 2007; Cherukuri et al., 2017; Liang and Stokes, 2019). This regularization has been widely used in both single-agent MDPs (Neu et al., 2017; Haarnoja et al., 2018; Chow et al., 2018; Geist et al., 2019), and learning in games (Syrgkanis et al., 2015; Mertikopoulos and Sandholm, 2016; Grill et al., 2019), to improve both the exploration and convergence.

With small enough τ_i (with the order of $\mathcal{O}(\epsilon)$, see §B.1), the solution to (6) will be ϵ-close to that of the unregularized one (Geist et al., 2019). More importantly, many commonly used regularizations, including negative entropy (Neu et al., 2017), Tsallis entropy (Chow et al., 2018) and Rényi entropy with certain parameters (Mertikopoulos and Sandholm, 2016), naturally yield a smooth **Planning Oracle**; see Lemma 24 in §B.1 for a formal statement. Note that the smoothness of the oracle does not affect the sample complexity of our model-based MARL algorithm.

Now we present another theorem, which gives the ϵ-Nash equilibrium *policy pair* directly, with the (near-)minimax optimal sample complexity of $\widetilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-3}\epsilon^{-2})$.

Theorem 8 (Finding ϵ-NE Policy with a Smooth Planning Oracle) *Suppose that the policy pair $(\widehat{\mu}, \widehat{\nu})$ is obtained from a smooth **Planning Oracle** using the empirical model $\widehat{\mathcal{G}}$ (see Definition 7), which satisfies*

$$\|\widehat{V}^{\widehat{\mu},*} - \widehat{V}^*\|_\infty \leq \epsilon_{opt}, \qquad \|\widehat{V}^{*,\widehat{\nu}} - \widehat{V}^*\|_\infty \leq \epsilon_{opt}.$$

Then, for any $\delta \in (0, 1]$ and $\epsilon \in (0, 1/(1-\gamma)^{1/2}]$, if

$$N \geq \frac{c\gamma \log\left[c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-4}\delta^{-1}\right]}{(1-\gamma)^3\epsilon^2}$$

for some absolute constant c, then, letting $\widetilde{\epsilon} := \epsilon + 4\epsilon_{opt}/(1-\gamma)$, with probability at least $1-\delta$,

$$V^{*,\widehat{\nu}} - 2\widetilde{\epsilon} \leq V^{\widehat{\mu},\widehat{\nu}} \leq V^{\widehat{\mu},*} + 2\widetilde{\epsilon},$$

namely, $(\widehat{\mu}, \widehat{\nu})$ constitutes a $2\widetilde{\epsilon}$-Nash equilibrium policy pair.

Theorem 8 shows that the sample complexity of achieving an ϵ-NE policy can be near-minimax optimal for the reward-aware case, and minimax-optimal for the reward-agnostic case, if a smooth **Planning Oracle** is used. The dependence on $|\mathcal{S}|$ and $1-\gamma$ also matches the only known near-optimal complexity in MGs in Sidford et al. (2020), with a turn-based setting and a model-free algorithm. Inherited from Agarwal et al. (2019a), this improves the second result in Azar et al. (2013) that also has $\widetilde{\mathcal{O}}((1-\gamma)^{-3}\epsilon^{-2})$ in finding an ϵ-optimal policy, by removing the dependence on $|\mathcal{S}|^{-1/2}$ and enlarging the choice of ϵ from $(0, (1-\gamma)^{-1/2}|\mathcal{S}|^{-1/2}]$ to $(0, (1-\gamma)^{-1/2}]$, and removing a factor of $|\mathcal{S}|$ in the total sample complexity for any fixed ϵ. In addition, Theorem 8 also applies to the multi-reward setting, as Theorem 5, by taking a union bound argument over all reward functions in the reward-agnostic case. If the number of reward functions M is of order poly$(|\mathcal{S}|, |\mathcal{A}|, |\mathcal{B}|)$, the sample complexity of handling multiple reward functions has the same order as that in Theorem 8.

Theorems 5 and 8 together justify that, this simple model-based MARL algorithm is indeed sample-efficient, in approximating both the Nash equilibrium values and policies.

Moreover, our separation of the reward-aware and reward-agnostic cases highlights both the power (easily handling multiple reward functions), and the limitation (less adaptive and can hardly achieve $\widetilde{\mathcal{O}}(|\mathcal{A}|+|\mathcal{B}|)$) of the model-based approach, particularly arising in the multi-agent RL context.

4. Proofs

We first introduce some additional notation for convenience.

Notation. For a matrix $X \in \mathbb{R}^{m\times n}$, $X \geq c$ for some scalar $c \in \mathbb{R}$ means that each element of X is no-less than c. For a vector x, we use (x^2), $\sqrt{x}$, $|x|$ to denote the component-wise square, square-root, and absolute value of x. We use $P_{(s,a,b),s'}$ to denote the transition probability $P(s' \,|\, s,a,b)$, and $P_{s,a,b}$ to denote the vector $P(\cdot \,|\, s,a,b)$. We also use $P^{\mu,\nu}$ to denote the transition probability of state-action pairs induced by the policy pair (μ,ν), which is defined as

$$P^{\mu,\nu}_{(s,a,b),(s',a',b')} = \mu(a' \,|\, s')\nu(b' \,|\, s')P(s' \,|\, s,a,b).$$

Hence, the Q-value function can be written as

$$Q^{\mu,\nu} = r + \gamma P^{\mu,\nu} Q^{\mu,\nu} = (I - \gamma P^{\mu,\nu})^{-1} r.$$

Also, for any $V \in \mathbb{R}^{|\mathcal{S}|}$, we define the vector $\mathrm{Var}_P(V) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{B}|}$ as

$$\mathrm{Var}_P(V)(s,a,b) := \mathrm{Var}_{P(\cdot \,|\, s,a,b)}(V) = P(V)^2 - (PV)^2.$$

Then, we define $\Sigma^{\mu,\nu}_{\mathcal{G}}$ to be the variance of the discounted reward under the MG $\mathcal{G}$, i.e.,

$$\Sigma^{\mu,\nu}_{\mathcal{G}}(s,a,b) := \mathbb{E}\Big[\Big(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,b_t) - Q^{\mu,\nu}_{\mathcal{G}}(s,a,b)\Big)^2 \,\Big|\, s_0 = s, a_0 = a, b_0 = b\Big].$$

It can be shown (see an almost identical formula for MDPs in (Azar et al., 2013, Lemma 6)) that $\Sigma^{\mu,\nu}_{\mathcal{G}}$ satisfies some *Bellman-type* equation for any policy pair (μ,ν):

$$\Sigma^{\mu,\nu}_{\mathcal{G}} = \gamma^2\,\mathrm{Var}_P(V^{\mu,\nu}_{\mathcal{G}}) + \gamma^2 P^{\mu,\nu}\Sigma^{\mu,\nu}_{\mathcal{G}}. \tag{7}$$

It can also be verified that $\|\Sigma^{\mu,\nu}_{\mathcal{G}}\|_\infty \leq \gamma^2/(1-\gamma)^2$ (Azar et al., 2013; Agarwal et al., 2019a). Before proceeding further, we provide a roadmap for the proof.

Proof Roadmap. Our proof mainly consists of the following steps:

1. **Helper lemmas and a crude bound.** We first establish several important lemmas, including the component-wise error bounds for the final Q-value errors, the variance error bound, and a crude error bound that directly uses Hoeffding's inequality. Some of the results are adapted from the single-agent setting to zero-sum MGs, see Agarwal et al. (2019a). See §4.1.

2. **Establishing an auxiliary Markov game.** To improve the crude bound, we build up an *absorbing Markov game*, in order to handle the statistical dependence between $\widehat{P}$ and some value function generated by $\widehat{P}$, which occurs as a product in the component-wise bound above. By carefully designing the auxiliary game, we establish a Bernstein-like concentration inequality, despite this dependency. See §4.2, more precisely, Lemmas 17 and 18.

3. **Final bound for ϵ-approximate NE value.** Lemma 17 in Step **2** allows us to exploit the variance bound, see Lemma 11, to obtain an $\widetilde{\mathcal{O}}(\sqrt{1/[(1-\gamma)^3]N})$ order bound on the Q-value error, leading to a $\widetilde{\mathcal{O}}((1-\gamma)^{-3}\epsilon^{-2})$ near-minimax optimal sample complexity for achieving the ϵ-approximate NE value. See §4.3.

4. **Final bounds for ϵ-NE policy.** Based on the final bound in Step **3**, we then establish a $\widetilde{\mathcal{O}}((1-\gamma)^{-5}\epsilon^{-2})$ sample complexity for obtaining an ϵ-NE policy pair, by solving an additional matrix game over the output Q-value $\widehat{Q}^{\widehat{\mu},\widehat{\nu}}$. See §4.4. In addition, given a smooth **Planning Oracle**, by Lemma 18 in Step **2**, and more careful self-bounding techniques, we establish a $\widetilde{\mathcal{O}}((1-\gamma)^{-3}\epsilon^{-2})$ sample complexity for achieving such an ϵ-NE policy, directly using the output policies $(\widehat{\mu},\widehat{\nu})$. See §4.5.

4.1 Important Lemmas

We start with the component-wise error bounds.

Lemma 9 (Component-Wise Bounds) *For any policy pair (μ,ν), it follows that*

$$Q^{\mu,\nu}-\widehat{Q}^{\mu,\nu}=\gamma(I-\gamma P^{\mu,\nu})^{-1}(P-\widehat{P})\widehat{V}^{\mu,\nu},$$
$$\gamma(I-\gamma\widehat{P}^{\mu,\nu(\mu)})^{-1}(P-\widehat{P})V^{\mu,*}\leq Q^{\mu,*}-\widehat{Q}^{\mu,*}\leq\gamma(I-\gamma P^{\mu,\widehat{\nu(\mu)}})^{-1}(P-\widehat{P})\widehat{V}^{\mu,*},$$
$$\gamma(I-\gamma P^{\widehat{\mu(\nu)},\nu})^{-1}(P-\widehat{P})\widehat{V}^{*,\nu}\leq Q^{*,\nu}-\widehat{Q}^{*,\nu}\leq\gamma(I-\gamma\widehat{P}^{\mu(\nu),\nu})^{-1}(P-\widehat{P})V^{*,\nu},$$

where we recall that $\nu(\mu)$ and $\mu(\nu)$ denote the best-response policy given μ and ν, respectively (see (4)). Moreover, we have

$$Q^{\mu,\nu}\geq Q^*-\|Q^{\mu,\nu}-\widehat{Q}^{\mu,\nu}\|_\infty-\|\widehat{Q}^{\mu,\nu}-\widehat{Q}^*\|_\infty-\|\widehat{Q}^{\mu^*,*}-Q^*\|_\infty \tag{8}$$
$$Q^{\mu,\nu}\leq Q^*+\|Q^{\mu,\nu}-\widehat{Q}^{\mu,\nu}\|_\infty+\|\widehat{Q}^{\mu,\nu}-\widehat{Q}^*\|_\infty+\|\widehat{Q}^{*,\nu^*}-Q^*\|_\infty \tag{9}$$
$$V^{\mu,*}\geq V^*-\|Q^{\mu,*}-\widehat{Q}^{\mu,*}\|_\infty-\|\widehat{V}^{\mu,*}-\widehat{V}^*\|_\infty-\|\widehat{Q}^{\mu^*,*}-Q^*\|_\infty \tag{10}$$
$$V^{*,\nu}\leq V^*+\|Q^{*,\nu}-\widehat{Q}^{*,\nu}\|_\infty+\|\widehat{V}^{*,\nu}-\widehat{V}^*\|_\infty+\|\widehat{Q}^{*,\nu^*}-Q^*\|_\infty. \tag{11}$$

Proof First, note that

$$Q^{\mu,\nu}-\widehat{Q}^{\mu,\nu}=(I-\gamma P^{\mu,\nu})^{-1}r-(I-\gamma\widehat{P}^{\mu,\nu})^{-1}r=(I-\gamma P^{\mu,\nu})^{-1}[(I-\gamma\widehat{P}^{\mu,\nu})-(I-\gamma P^{\mu,\nu})]\widehat{Q}^{\mu,\nu}$$
$$=\gamma(I-\gamma P^{\mu,\nu})^{-1}(P^{\mu,\nu}-\widehat{P}^{\mu,\nu})\widehat{Q}^{\mu,\nu}=\gamma(I-\gamma P^{\mu,\nu})^{-1}(P-\widehat{P})\widehat{V}^{\mu,\nu},$$

proving the first equation. Also,

$$Q^{\mu,*}-\widehat{Q}^{\mu,*}\leq Q^{\mu,\widehat{\nu(\mu)}}-\widehat{Q}^{\mu,*}=Q^{\mu,\widehat{\nu(\mu)}}-\widehat{Q}^{\mu,\widehat{\nu(\mu)}}$$
$$=\big(I-\gamma P^{\mu,\widehat{\nu(\mu)}}\big)^{-1}r-\big(I-\gamma\widehat{P}^{\mu,\widehat{\nu(\mu)}}\big)^{-1}r$$
$$=\big(I-\gamma P^{\mu,\widehat{\nu(\mu)}}\big)^{-1}\big[(I-\gamma\widehat{P}^{\mu,\widehat{\nu(\mu)}})-(I-\gamma P^{\mu,\widehat{\nu(\mu)}})\big]\widehat{Q}^{\mu,\widehat{\nu(\mu)}}=\gamma(I-\gamma P^{\mu,\widehat{\nu(\mu)}})^{-1}(P-\widehat{P})\widehat{V}^{\mu,\widehat{\nu(\mu)}},$$

where we recall that $\widehat{\nu(\mu)}(\cdot\,|\,s) \in \operatorname{argmin} \widehat{V}^{\mu,\nu}(s)$ for all $s \in \mathcal{S}$. By similar arguments, recalling that $\nu(\mu)(\cdot\,|\,s) \in \operatorname{argmin} V^{\mu,\nu}(s)$ for all s, we have

$$Q^{\mu,*} - \widehat{Q}^{\mu,*} \geq Q^{\mu,\nu(\mu)} - \widehat{Q}^{\mu,\nu(\mu)} = (I - \gamma P^{\mu,\nu(\mu)})^{-1} r - (I - \gamma \widehat{P}^{\mu,\nu(\mu)})^{-1} r$$
$$= (I - \gamma \widehat{P}^{\mu,\nu(\mu)})^{-1} [(I - \gamma \widehat{P}^{\mu,\nu(\mu)}) - (I - \gamma P^{\mu,\nu(\mu)})] Q^{\mu,\nu(\mu)} = \gamma (I - \gamma \widehat{P}^{\mu,\nu(\mu)})^{-1} (P - \widehat{P}) V^{\mu,*}.$$

Similar arguments yield the third inequality in the first argument.

For the second argument, we have

$$Q^{\mu,\nu} - Q^* = Q^{\mu,\nu} - \widehat{Q}^* + \widehat{Q}^* - Q^* \geq Q^{\mu,\nu} - \widehat{Q}^* + \widehat{Q}^{\mu^*,*} - Q^* \geq -\|Q^{\mu,\nu} - \widehat{Q}^*\|_\infty - \|\widehat{Q}^{\mu^*,*} - Q^*\|_\infty,$$

which, combined with triangle inequality, yields the first inequality. Similarly, we have

$$Q^{\mu,\nu} - Q^* = Q^{\mu,\nu} - \widehat{Q}^* + \widehat{Q}^* - Q^* \leq Q^{\mu,\nu} - \widehat{Q}^* + \widehat{Q}^{*,\nu^*} - Q^* \leq \|Q^{\mu,\nu} - \widehat{Q}^*\|_\infty + \|\widehat{Q}^{*,\nu^*} - Q^*\|_\infty.$$

Using triangle inequality proves the second inequality. For (10)-(11), we similarly have

$$V^{\mu,*} - V^* = V^{\mu,*} - \widehat{V}^* + \widehat{V}^* - V^* \geq V^{\mu,*} - \widehat{V}^* + \widehat{V}^{\mu^*,*} - V^* \geq -\|V^{\mu,*} - \widehat{V}^*\|_\infty - \|\widehat{V}^{\mu^*,*} - V^*\|_\infty, \tag{12}$$

$$V^{*,\nu} - V^* = V^{*,\nu} - \widehat{V}^* + \widehat{V}^* - V^* \leq V^{*,\nu} - \widehat{V}^* + \widehat{V}^{*,\nu} - V^* \leq \|V^{*,\nu} - \widehat{V}^*\|_\infty + \|\widehat{V}^{*,\nu} - V^*\|_\infty. \tag{13}$$

Notice that for any $\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}$ and $\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}$,

$$\|V^{\mu,*} - \widehat{V}^{\mu,*}\|_\infty = \Big\| \min_{\vartheta \in \Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu(\cdot\,|\,s), b\sim\vartheta}[Q^{\mu,*}(\cdot,a,b)] - \min_{\vartheta \in \Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu(\cdot\,|\,s), b\sim\vartheta}[\widehat{Q}^{\mu,*}(\cdot,a,b)] \Big\|_\infty$$
$$\leq \max_{\vartheta \in \Delta(\mathcal{B})} \big\| \mathbb{E}_{a\sim\mu(\cdot\,|\,s), b\sim\vartheta}[Q^{\mu,*}(\cdot,a,b)] - \mathbb{E}_{a\sim\mu(\cdot\,|\,s), b\sim\vartheta}[\widehat{Q}^{\mu,*}(\cdot,a,b)] \big\|_\infty \leq \|Q^{\mu,*} - \widehat{Q}^{\mu,*}\|_\infty \tag{14}$$

$$\|V^{*,\nu} - \widehat{V}^{*,\nu}\|_\infty = \Big\| \max_{u \in \Delta(\mathcal{A})} \mathbb{E}_{a\sim u, b\sim\nu(\cdot\,|\,s)}[Q^{*,\nu}(\cdot,a,b)] - \max_{u \in \Delta(\mathcal{A})} \mathbb{E}_{a\sim u, b\sim\nu(\cdot\,|\,s)}[\widehat{Q}^{*,\nu}(\cdot,a,b)] \Big\|_\infty$$
$$\leq \max_{u \in \Delta(\mathcal{A})} \big\| \mathbb{E}_{a\sim u, b\sim\nu(\cdot\,|\,s)}[Q^{*,\nu}(\cdot,a,b)] - \mathbb{E}_{a\sim u, b\sim\nu(\cdot\,|\,s)}[\widehat{Q}^{*,\nu}(\cdot,a,b)] \big\|_\infty \leq \|Q^{*,\nu} - \widehat{Q}^{*,\nu}\|_\infty. \tag{15}$$

Combining (12)-(13) and (14)-(15), together with triangle inequality, we arrive at (10)-(11), and complete the proof. ■

We establish the decomposition in (8)-(9) for the following intuition and reasons. The error in (8)-(9) contains three terms: the first and third terms $\|Q^{\mu,\nu} - \widehat{Q}^{\mu,\nu}\|_\infty$ and $\widehat{Q}^*\|_\infty - \|\widehat{Q}^{\mu^*,*} - Q^*\|_\infty$ are the differences of the Q-value for some policy pairs in the true and estimated models, respectively, which will be handled later based on the *statistical error* of the model estimation; the second term $\|\widehat{Q}^{\mu,\nu} - \widehat{Q}^*\|_\infty$ is the *optimization error* we obtained from the algorithm that solves the empirical game, which will be controlled with an efficient **Planning Oracle**. To deal with the statistical errors, we first introduce the following lemma, which is adapted from Lemma 2 in Agarwal et al. (2019a).

Lemma 10 *For any policy pair (μ,ν) and vector $v \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{B}|}$, $\|(I - \gamma P^{\mu,\nu})^{-1} v\|_\infty \leq \|v\|_\infty / (1-\gamma)$.*

Proof The proof is straightforward. Letting $w = (I-\gamma P^{\mu,\nu})^{-1}v$, we have $v = (I-\gamma P^{\mu,\nu})w$. Triangle inequality yields $\|v\|_\infty \geq \|w\|_\infty - \gamma\|P^{\mu,\nu}w\|_\infty \geq \|w\|_\infty - \gamma\|w\|_\infty$, which completes the proof. ■

Next we establish the Bellman property of a policy pair (μ,ν)'s variance and its accumulation. This has been observed for MDPs before in Munos and Moore (1999); Lattimore and Hutter (2012); Azar et al. (2012); Agarwal et al. (2019a). We establish the counterpart for Markov games as follows.

Lemma 11 *For any policy pair (μ,ν) and MG $\mathcal{G}$ with transition model P, we have*

$$\left\|\left(I-\gamma P^{\mu,\nu}\right)^{-1}\sqrt{\mathrm{Var}_P\left(V_{\mathcal{G}}^{\mu,\nu}\right)}\right\|_\infty \leq \sqrt{\frac{2}{(1-\gamma)^3}}.$$

Proof The proof follows that of (Agarwal et al., 2019a, Lemma 3). For any positive vector v, by Jensen's inequality, we have

$$\|(I-\gamma P^{\mu,\nu})^{-1}\sqrt{v}\|_\infty = \frac{1}{1-\gamma}\|(1-\gamma)(I-\gamma P^{\mu,\nu})^{-1}\sqrt{v}\|_\infty \leq \sqrt{\left\|\frac{1}{1-\gamma}(I-\gamma P^{\mu,\nu})^{-1}v\right\|_\infty}. \tag{16}$$

Also, observe that

$$\begin{aligned}
&\|(I-\gamma P^{\mu,\nu})^{-1}v\|_\infty = \|(I-\gamma P^{\mu,\nu})^{-1}(I-\gamma^2 P^{\mu,\nu})(I-\gamma^2 P^{\mu,\nu})^{-1}v\|_\infty\\
&\quad= \left\|[(I-\gamma P^{\mu,\nu})^{-1}(1-\gamma+\gamma-\gamma^2 P^{\mu,\nu})](I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty\\
&\quad= \left\|[(1-\gamma)(I-\gamma P^{\mu,\nu})^{-1}+\gamma I](I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty\\
&\quad\leq (1-\gamma)\left\|(I-\gamma P^{\mu,\nu})^{-1}(I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty + \gamma\left\|(I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty\\
&\quad\leq \frac{1-\gamma}{1-\gamma}\left\|(I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty + \gamma\left\|(I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty \leq 2\left\|(I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty.
\end{aligned} \tag{17}$$

Combining (16) and (17) yields

$$\|(I-\gamma P^{\mu,\nu})^{-1}\sqrt{v}\|_\infty \leq \sqrt{\left\|\frac{2}{1-\gamma}(I-\gamma^2 P^{\mu,\nu})^{-1}v\right\|_\infty}. \tag{18}$$

In addition, by (7), we have $\Sigma_{\mathcal{G}}^{\mu,\nu} = \gamma^2(I-\gamma^2 P^{\mu,\nu})^{-1}\,\mathrm{Var}_P(V_{\mathcal{G}}^{\mu,\nu})$. Letting $v = \mathrm{Var}_P(V_{\mathcal{G}}^{\mu,\nu})$ in (18) and noticing that $\|\Sigma_{\mathcal{G}}^{\mu,\nu}\|_\infty \leq \gamma^2/(1-\gamma)^2$ completes the proof. ■

Finally, if we just apply Hoeffding's inequality, we obtain the following concentration argument, upon which we will improve to obtain our final results.

Lemma 12 *Let (μ^*,ν^*) be the Nash equilibrium policy pair under the actual model $\mathcal{G}$. Then, for any $\delta\in(0,1]$, with probability at least $1-\delta$, we have*

$$\|Q^*-\widehat{Q}^{\mu^*,\nu^*}\|_\infty \leq \Delta_{\delta,N},\quad \|Q^*-\widehat{Q}^{\mu^*,*}\|_\infty \leq \Delta_{\delta,N},\quad \|Q^*-\widehat{Q}^{*,\nu^*}\|_\infty \leq \Delta_{\delta,N},\quad \|Q^*-\widehat{Q}^{*}\|_\infty \leq \Delta_{\delta,N},$$

where

$$\Delta_{\delta,N} := \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}.$$

Proof First note that V^* is fixed and independent of the randomness in $\widehat{P}$. Due to the boundedness of V^* that $\|V^*\|_\infty \leq (1-\gamma)^{-1}$, and the union of Hoeffding bounds over $\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, we have that with probability at least $1-\delta$

$$\big\|(\widehat{P}-P)V^*\big\|_\infty \leq \frac{1}{1-\gamma}\cdot\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}. \tag{19}$$

On the other hand, let $\mathcal{T}_{\mu,\nu}$ be the Bellman operator under the true transition model P, using any joint policy (μ,ν), i.e., for any $s\in\mathcal{S}$ and $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, $V\in\mathbb{R}^{|\mathcal{S}|}$ and $Q\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{B}|}$:

$$\begin{aligned}\mathcal{T}_{\mu,\nu}(V)(s) &= \mathbb{E}_{a\sim\mu(\cdot\,|\,s),b\sim\nu(\cdot\,|\,s)}\big[r(s,a,b)+\gamma\cdot P(\cdot\,|\,s,a,b)^\top V\big]\\ \mathcal{T}_{\mu,\nu}(Q)(s,a,b) &= r(s,a,b)+\gamma\cdot\mathbb{E}_{s'\sim P(\cdot\,|\,s,a,b),a'\sim\mu(\cdot\,|\,s'),b'\sim\nu(\cdot\,|\,s')}\big[Q(s',a',b')\big].\end{aligned}$$

Similarly, let $\widehat{\mathcal{T}}_{\mu,\nu}$ be the corresponding operator defined under the estimated transition $\widehat{P}$. Note that $\widehat{Q}^{\mu,\nu}$ and Q^* are the fixed points of $\widehat{\mathcal{T}}_{\mu,\nu}$ and $\mathcal{T}_{\mu^*,\nu^*}$, respectively. We thus have

$$\begin{aligned}&\|Q^*-\widehat{Q}^{\mu,\nu}\|_\infty = \|\mathcal{T}_{\mu^*,\nu^*}Q^*-\widehat{\mathcal{T}}_{\mu,\nu}\widehat{Q}^{\mu,\nu}\|_\infty\\ &\quad\leq \|\mathcal{T}_{\mu^*,\nu^*}Q^*-r-\gamma\widehat{P}^{\mu^*,\nu^*}Q^*\|_\infty+\|r+\gamma\widehat{P}^{\mu^*,\nu^*}Q^*-\widehat{\mathcal{T}}_{\mu,\nu}\widehat{Q}^{\mu,\nu}\|_\infty\\ &\quad= \gamma\|P^{\mu^*,\nu^*}Q^*-\widehat{P}^{\mu^*,\nu^*}Q^*\|_\infty+\gamma\|\widehat{P}^{\mu^*,\nu^*}Q^*-\widehat{P}^{\mu,\nu}\widehat{Q}^{\mu,\nu}\|_\infty\\ &\quad= \gamma\|PV^*-\widehat{P}V^*\|_\infty+\gamma\|\widehat{P}V^*-\widehat{P}\widehat{V}^{\mu,\nu}\|_\infty \leq \gamma\|(P-\widehat{P})V^*\|_\infty+\gamma\|V^*-\widehat{V}^{\mu,\nu}\|_\infty.\end{aligned} \tag{20}$$

To show the first argument, letting $\mu=\mu^*$ and $\nu=\nu^*$, we have

$$\begin{aligned}\gamma\|V^*-\widehat{V}^{\mu^*,\nu^*}\|_\infty &= \gamma\big\|\mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\nu^*(\cdot\,|\,s)}[Q^*(\cdot,a,b)]-\mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\nu^*(\cdot\,|\,s)}[\widehat{Q}^{\mu^*,\nu^*}(\cdot,a,b)]\big\|_\infty\\ &\leq \gamma\|Q^*-\widehat{Q}^{\mu^*,\nu^*}\|_\infty.\end{aligned} \tag{21}$$

Using (21) to bound the last term in (20), and solving for $\|Q^*-\widehat{Q}^{\mu^*,\nu^*}\|_\infty$ from (20), we obtain the first argument.

For the second argument, letting $\mu=\mu^*$ and $\nu=\widehat{\nu(\mu^*)}$ (note that $\widehat{Q}^{\mu^*,*}=\widehat{Q}^{\mu^*,\widehat{\nu(\mu^*)}}$), we have

$$\begin{aligned}\gamma\|V^*-\widehat{V}^{\mu^*,*}\|_\infty &= \gamma\cdot\big\|\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}[Q^*(\cdot,a,b)]-\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}[\widehat{Q}^{\mu^*,*}(\cdot,a,b)]\big\|_\infty\\ &\leq \gamma\cdot\max_{\vartheta\in\Delta(\mathcal{B})}\big\|\mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}[Q^*(\cdot,a,b)]-\mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}[\widehat{Q}^{\mu^*,*}(\cdot,a,b)]\big\|_\infty \leq \gamma\|Q^*-\widehat{Q}^{\mu^*,*}\|_\infty,\end{aligned} \tag{22}$$

where the first inequality is due to the non-expansiveness of the min operator. Using (22) to bound the last term in (20), and solving for $\|Q^*-\widehat{Q}^{\mu^*,*}\|_\infty$ from (20), we obtain the second argument. Similarly, we can obtain the third argument.

For the fourth argument, letting $\mu=\widehat{\mu}^*$ and $\nu=\widehat{\nu}^*$, the NE policy under $\widehat{P}$ (note that $\widehat{Q}^{\widehat{\mu}^*,\widehat{\nu}^*}=\widehat{Q}^*$), we have

$$\begin{aligned}\gamma\|V^*-\widehat{V}^*\|_\infty &= \gamma\cdot\big\|\max_{u\in\Delta(\mathcal{A})}\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}[Q^*(\cdot,a,b)]-\max_{u\in\Delta(\mathcal{A})}\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}[\widehat{Q}^*(\cdot,a,b)]\big\|_\infty\\ &\leq \gamma\cdot\max_{u\in\Delta(\mathcal{A})}\big\|\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}[Q^*(\cdot,a,b)]-\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}[\widehat{Q}^*(\cdot,a,b)]\big\|_\infty \leq \gamma\|Q^*-\widehat{Q}^*\|_\infty,\end{aligned}$$

where the inequalities are due to the non-expansivenesses of both the max and the min operators. This, combined with (20), completes the proof. ■

The argument above will lead to a crude bound, with an additional $1/(1-\gamma)$ dependence compared to our main results in Theorem 5 and Theorem 8. The key reason is that we used some self-bounding of the error terms, e.g., $\|Q^* - \widehat{Q}^{\mu^*,\nu^*}\|_\infty$, which appears on both sides of the inequality, with a γ discounting on the right-hand side. This way, by subtracting the term on the right-hand side, we have an additional $1/(1-\gamma)$ order after dividing $(1-\gamma)$ on both sides. This was essentially due to the fact that the direct concentration argument can only deal with the concentration of $\big\|(\widehat{P}-P)V^*\big\|_\infty$, where $\widehat{P}$ and V^* are *not* dependent as V^* is a fixed vector. To obtain sharper rates, one has to directly deal with the quantities as $\big\|(\widehat{P}-P)\widehat{V}\big\|_\infty$, where $\widehat{V}$ denotes some value function obtained from the empirical model, and is correlated with $\widehat{P}$. Properly handling this interdependence will be the focus of our proof next.

4.2 An Auxiliary Markov Game

Motivated by the *absorbing MDP* technique in Agarwal et al. (2019a), we introduce an *absorbing Markov game*, in order to handle the interdependence between $\widehat{P}$ and $\widehat{V}^{\mu,\nu}$, for any μ, ν (which may also depend on $\widehat{P}$), which will show up frequently in the analysis.

We now define a new Markov game $\mathcal{G}_{s,u}$ as follows (with $s \in \mathcal{S}$ and $u \in \mathbb{R}$ a constant): $\mathcal{G}_{s,u}$ is identical to $\mathcal{G}$, except that $P_{\mathcal{G}_{s,u}}(s \,|\, s,a,b) = 1$ for all $(a,b) \in \mathcal{A}\times\mathcal{B}$, namely, state s is an *absorbing* state; and the instantaneous reward at s is always $(1-\gamma)u$. The rest of the reward function and the transition model of $\mathcal{G}_{s,u}$ are the same as those of $\mathcal{G}$. For notational simplicity, we now use $X^{\mu,\nu}_{s,u}$ to denote $X^{\mu,\nu}_{\mathcal{G}_{s,u}}$, where X can be either the value functions Q and V, or the reward function r, under the model $\mathcal{G}_{s,u}$. Obviously, for any policy pair (μ,ν), $V^{\mu,\nu}_{s,u}(s) = u$ for the absorbing state s.

In addition, we define U_s for some state s to choose u from, which is a set of evenly spaced elements in the interval $[V^*(s)-\Delta, V^*(s)+\Delta]$ for some $\Delta > 0$, i.e., $U_s \subset [V^*(s)-\Delta, V^*(s)+\Delta]$. An appropriately chosen size of $|U_s|$ will be the key in the proof. We also use $\widehat{P}_{\mathcal{G}_{s,u}}$ to denote the transition model of the absorbing MG for the empirical MG $\widehat{\mathcal{G}}$, denoted by $\widehat{\mathcal{G}}_{s,u}$. Specifically, at all non-absorbing states, $\widehat{P}_{\mathcal{G}_{s,u}}$ is identical to $\widehat{P}$; while at the absorbing state, $\widehat{P}_{\mathcal{G}_{s,u}}(s \,|\, s,a,b) = 1$ for any $(a,b) \in \mathcal{A}\times\mathcal{B}$. The corresponding value functions are for short denoted by $\widehat{V}^{\mu,\nu}_{s,u}$ and $\widehat{Q}^{\mu,\nu}_{s,u}$. Similar as in the original MG, we also use $\widehat{V}^*_{s,u}$ to denote the NE value under the model $\widehat{\mathcal{G}}_{s,u}$, and use $\widehat{V}^{\mu,*}_{s,u}$ and $\widehat{V}^{*,\nu}_{s,u}$ to denote the best-response values of some given μ and ν, under the model $\widehat{\mathcal{G}}_{s,u}$. Now we first have the following lemma based on Bernstein's inequality; see a similar argument in Lemma 5 in Agarwal et al. (2019a).

Lemma 13 *For fixed state s, action (a,b), a finite set U_s, and $\delta > 0$, it holds that for all $u \in U_s$, with probability greater than $1-\delta$,*

$$\big|(P_{s,a,b} - \widehat{P}_{s,a,b}) \cdot \widehat{V}^*_{s,u}\big| \leq \sqrt{\frac{2\log(4|U_s|/\delta) \cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^*_{s,u})}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N},$$

$$\big|(P_{s,a,b} - \widehat{P}_{s,a,b}) \cdot \widehat{V}^{\mu^*,*}_{s,u}\big| \leq \sqrt{\frac{2\log(4|U_s|/\delta) \cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^{\mu^*,*}_{s,u})}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N},$$

$$\big|(P_{s,a,b} - \widehat{P}_{s,a,b}) \cdot \widehat{V}^{*,\nu^*}_{s,u}\big| \leq \sqrt{\frac{2\log(4|U_s|/\delta) \cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^{*,\nu^*}_{s,u})}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N},$$

$$\big|(P_{s,a,b} - \widehat{P}_{s,a,b}) \cdot \widehat{V}^{\mu^*,\nu^*}_{s,u}\big| \leq \sqrt{\frac{2\log(4|U_s|/\delta) \cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^{\mu^*,\nu^*}_{s,u})}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N},$$

$$\big|(P_{s,a,b} - \widehat{P}_{s,a,b}) \cdot V^{\widehat{\mu}_{s,u},*}\big| \leq \sqrt{\frac{2\log(4|U_s|/\delta) \cdot \mathrm{Var}_{P_{s,a,b}}(V^{\widehat{\mu}_{s,u},*})}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N},$$

$$\big|(P_{s,a,b} - \widehat{P}_{s,a,b}) \cdot V^{*,\widehat{\nu}_{s,u}}\big| \leq \sqrt{\frac{2\log(4|U_s|/\delta) \cdot \mathrm{Var}_{P_{s,a,b}}(V^{*,\widehat{\nu}_{s,u}})}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N},$$

where $P_{s,a,b}$ and $\widehat{P}_{s,a,b}$ are the transition models extracted from the original game $\mathcal{G}$ and its empirical version $\widehat{\mathcal{G}}$, respectively (not related to either $\mathcal{G}_{s,u}$ or $\widehat{\mathcal{G}}_{s,u}$), and $(\widehat{\mu}_{s,a}, \widehat{\nu}_{s,a})$ is the output of the ***Planning Oracle*** *using the auxiliary empirical model $\widehat{\mathcal{G}}_{s,u}$*

Proof The key observation is that the random variables $\widehat{P}_{s,a,b}$ and $\widehat{V}^*_{s,u}$ are independent. Using Bernstein's inequality along with a union bound over all $u \in U_s$, we obtain the first inequality. The other inequalities follow similarly, as $\widehat{P}_{s,a,b}$ is independent of $\widehat{V}^{\mu^*,*}_{s,u}$, $\widehat{V}^{*,\nu^*}_{s,u}$, $\widehat{V}^{\mu^*,\nu^*}_{s,u}$, $V^{\widehat{\mu}_{s,u},*}$, and $V^{*,\widehat{\nu}_{s,u}}$. This is because the latter terms are all decided by the original game $\mathcal{G}$, and/or the auxiliary empirical game $\widehat{\mathcal{G}}_{s,u}$ (not the original empirical game $\widehat{\mathcal{G}}$). ■

Note that the arguments in Lemma 13 do not hold, if we replace $\widehat{V}^*_{s,u}$ by $\widehat{V}^*$, or $\widehat{V}^{\mu^*,*}_{s,u}$ by $\widehat{V}^{\mu^*,*}$, or $\widehat{V}^{*,\nu^*}_{s,u}$ by $\widehat{V}^{*,\nu^*}$. It will neither hold if we replace $\widehat{V}^{\mu^*,*}_{s,u}$ and $V^{\widehat{\mu}_{s,u},*}$ by some $\widehat{V}^{\mu,*}$ and $V^{\mu,*}$, for any μ that is dependent on $\widehat{P}$, e.g., the NE policy $\widehat{\mu}^*$ for the original empirical game $\widehat{\mathcal{G}}$. This is one of the key subtleties that are worth emphasizing.

Next we establish two helpful lemmas that help guide the choices of U_s, so that $\widehat{V}^*_{s,u}$ (resp. $\widehat{V}^{\mu^*,*}_{s,u}$, $\widehat{V}^{*,\nu^*}_{s,u}$, and $\widehat{V}^{\mu^*,\nu^*}_{s,u}$) will be a good approximate of $\widehat{V}^*$ (resp. $\widehat{V}^{\mu^*,*}$, $\widehat{V}^{*,\nu^*}$, and $\widehat{V}^{\mu^*,\nu^*}$).

Lemma 14 *For the absorbing state s, and any joint policy (μ,ν), suppose that $u^* = V^*_{\mathcal{G}}(s)$, $u^{\mu,*} = V^{\mu,*}_{\mathcal{G}}(s)$, $u^{*,\nu} = V^{*,\nu}_{\mathcal{G}}(s)$, and $u^{\mu,\nu} = V^{\mu,\nu}_{\mathcal{G}}(s)$. Then,*

$$V^*_{\mathcal{G}} = V^*_{s,u^*} \qquad V^{\mu,*}_{\mathcal{G}} = V^{\mu,*}_{\mathcal{G}_{s,u^{\mu,*}}} \qquad V^{*,\nu}_{\mathcal{G}} = V^{*,\nu}_{\mathcal{G}_{s,u^{*,\nu}}} \qquad V^{\mu,\nu}_{\mathcal{G}} = V^{\mu,\nu}_{\mathcal{G}_{s,u^{\mu,\nu}}}.$$

Proof For the first formula, we need to verify that $V^*_{\mathcal{G}}$ satisfies the optimal (Nash equilibrium) Bellman equation for the game $\mathcal{G}_{s,u^*}$. To this end, note that if $s'=s$, then $u^*=V^*_{\mathcal{G}}(s)$ satisfies the Bellman equation trivially, since s is absorbing with the value $V^*_{s,u^*}(s)=u^*$.

On the other hand, for any $s'\neq s$, the outgoing transition model at s' in $\mathcal{G}_{s,u^*}$ is the same as that in $\mathcal{G}$, and $V^*_{\mathcal{G}}(s')$ per se satisfies the Bellman equation in $\mathcal{G}$ (which are the same for $\mathcal{G}_{s,u^*}$ at these states $s'\neq s$). Thus, $V^*_{\mathcal{G}}$ satisfies the Bellman equation in $\mathcal{G}_{s,u^*}$ for all states. This proves the first equation. The proofs for the remaining three equations are analogous. ■

Perfect choices of u have been specified in Lemma 14 above. Moreover, we need to quantify how the value changes if we deviate from these perfect choices, i.e., the robustness to misspecification of u (Agarwal et al., 2019a). This result is formally established in the following lemma; see also Lemma 7 in Agarwal et al. (2019a) for a similar result.

Lemma 15 *For any state s, $u,u'\in\mathbb{R}$, and joint policy pair (μ,ν), we have*

$$\big\|V^*_{s,u}-V^*_{s,u'}\big\|_\infty\le|u-u'|,\qquad \big\|V^{\mu,*}_{s,u}-V^{\mu,*}_{s,u'}\big\|_\infty\le|u-u'|,$$
$$\big\|V^{*,\nu}_{s,u}-V^{*,\nu}_{s,u'}\big\|_\infty\le|u-u'|,\qquad \big\|V^{\mu,\nu}_{s,u}-V^{\mu,\nu}_{s,u'}\big\|_\infty\le|u-u'|.$$

Proof Note that $\|r_{s,u}-r_{s,u'}\|_\infty=(1-\gamma)|u-u'|$, since the reward functions only differ at s, where $r_{s,u}(s,a,b)=(1-\gamma)u$ and $r_{s,u'}(s,a,b)=(1-\gamma)u'$. We denote the NE policy pair in $\mathcal{G}_{s,u}$ by $(\mu^*_{s,u},\nu^*_{s,u})$. Thus,

$$Q^*_{s,u}-Q^*_{s,u'}=Q^{\mu^*_{s,u},\nu^*_{s,u}}_{s,u}-Q^{\mu^*_{s,u'},\nu^*_{s,u'}}_{s,u'}\le Q^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u}-Q^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u'} \tag{23}$$
$$=\big(I-\gamma P^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u}\big)^{-1}r_{s,u}-\big(I-\gamma P^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u'}\big)^{-1}r_{s,u'} \tag{24}$$
$$=\big(I-\gamma P^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u}\big)^{-1}\big(r_{s,u}-r_{s,u'}\big) \tag{25}$$
$$\le\frac{\|r_{s,u}-r_{s,u'}\|_\infty}{1-\gamma}=|u-u'|, \tag{26}$$

where (23) uses the fact that at the NE,

$$V^{\mu^*_{s,u},\nu^*_{s,u}}_{s,u}=\min_\nu V^{\mu^*_{s,u},\nu}_{s,u}\le V^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u},\quad V^{\mu^*_{s,u'},\nu^*_{s,u'}}_{s,u'}=\max_\mu V^{\mu,\nu^*_{s,u'}}_{s,u'}\ge V^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u'},$$

implying the relationships of the corresponding Q-values; (24) is by definition; (25) uses the observation that $P^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u}$ is the same as $P^{\mu^*_{s,u},\nu^*_{s,u'}}_{s,u'}$ (transition is not affected by the value of u). Similarly, we can establish the lower bound that $Q^*_{s,u}-Q^*_{s,u'}\ge-|u-u'|$, which proves $\|Q^*_{s,u}-Q^*_{s,u'}\|_\infty\le|u-u'|$. Moreover, we have

$$\big\|V^*_{s,u}-V^*_{s,u'}\big\|_\infty=\Big\|\max_{u\in\Delta(\mathcal{A})}\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}[Q^*_{s,u}(\cdot,a,b)]-\max_{u\in\Delta(\mathcal{A})}\min_{\vartheta\in\Delta(\mathcal{B})}\mathbb{E}_{a\sim u,b\sim\vartheta}[Q^*_{s,u'}(\cdot,a,b)]\Big\|_\infty$$
$$\le\max_{u\in\Delta(\mathcal{A}),\vartheta\in\Delta(\mathcal{B})}\Big\|\mathbb{E}_{a\sim\mu(\cdot\,|\,s),b\sim\vartheta}[Q^*_{s,u}(\cdot,a,b)]-\mathbb{E}_{a\sim\mu(\cdot\,|\,s),b\sim\vartheta}[Q^*_{s,u'}(\cdot,a,b)]\Big\|_\infty$$
$$\le\big\|Q^*_{s,u}-Q^*_{s,u'}\big\|_\infty\le|u-u'|,$$

which proves the first inequality.

For the second one, recalling that the best-response policy of μ under $\mathcal{G}_{s,u}$ being $\nu_{s,u}(\mu)$, we have

$$
\begin{aligned}
Q^{\mu,*}_{s,u} - Q^{\mu,*}_{s,u'} &= \min_{\nu} Q^{\mu,\nu}_{s,u} - Q^{\mu,*}_{s,u'} = \min_{\nu} \big(I - \gamma P^{\mu,\nu}_{s,u}\big)^{-1} r_{s,u} - Q^{\mu,*}_{s,u'} && (27)\\
&\le \big(I - \gamma P^{\mu,\nu_{s,u'}(\mu)}_{s,u}\big)^{-1} r_{s,u} - \big(I - \gamma P^{\mu,\nu_{s,u'}(\mu)}_{s,u'}\big)^{-1} r_{s,u'} && (28)\\
&= \big(I - \gamma P^{\mu,\nu_{s,u'}(\mu)}_{s,u}\big)^{-1} \big(r_{s,u} - r_{s,u'}\big) \le \frac{\|r_{s,u} - r_{s,u'}\|_\infty}{1-\gamma} = |u - u'|, && (29)
\end{aligned}
$$

where (27) uses the definition of a best-response value, (28) plugs in the best-response policy $\nu_{s,u'}(\mu)$, and (29) also uses the fact that the transition does not depend on the value u. A lower bound can be established by noticing that $Q^{\mu,*}_{s,u'} = \min_\nu Q^{\mu,\nu}_{s,u'} \le Q^{\mu,\nu_{s,u}(\mu)}_{s,u'}$. This proves $\|Q^{\mu,*}_{s,u} - Q^{\mu,*}_{s,u'}\|_\infty \le |u - u'|$. Furthermore, notice that

$$
\begin{aligned}
\big\|V^{\mu,*}_{s,u} - V^{\mu,*}_{s,u'}\big\|_\infty &= \Big\| \min_{\vartheta \in \Delta(\mathcal{B})} \mathbb{E}_{a \sim \mu(\cdot \,|\, s), b \sim \vartheta}[Q^{\mu,*}_{s,u}(\cdot, a, b)] - \min_{\vartheta \in \Delta(\mathcal{B})} \mathbb{E}_{a \sim \mu(\cdot \,|\, s), b \sim \vartheta}[Q^{\mu,*}_{s,u'}(\cdot, a, b)] \Big\|_\infty \\
&\le \max_{\vartheta \in \Delta(\mathcal{B})} \Big\| \mathbb{E}_{a \sim \mu(\cdot \,|\, s), b \sim \vartheta}[Q^{\mu,*}_{s,u}(\cdot, a, b)] - \mathbb{E}_{a \sim \mu(\cdot \,|\, s), b \sim \vartheta}[Q^{\mu,*}_{s,u'}(\cdot, a, b)] \Big\|_\infty \\
&\le \big\|Q^{\mu,*}_{s,u} - Q^{\mu,*}_{s,u'}\big\|_\infty \le |u - u'|,
\end{aligned}
$$

which proves the second inequality. Similar arguments can also be used to establish the third and the fourth inequalities. This completes the proof. ■

We are now ready to show the main result in this section.

Lemma 16 *For any state s, joint action pair (a,b), and a finite set U_s, with probability greater than $1-\delta$, we have*

$$
\begin{aligned}
\big|(P_{s,a,b}-\widehat{P}_{s,a,b})\widehat{V}^*\big| \le& \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^*)}{N}}+\frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N}\\
&+\min_{u\in U_s}\big|\widehat{V}^*(s)-u\big|\cdot\left(2+\sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right)\\
\big|(P_{s,a,b}-\widehat{P}_{s,a,b})\widehat{V}^{\mu^*,*}\big| \le& \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^{\mu^*,*})}{N}}+\frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N}\\
&+\min_{u\in U_s}\big|\widehat{V}^{\mu^*,*}(s)-u\big|\cdot\left(2+\sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right)\\
\big|(P_{s,a,b}-\widehat{P}_{s,a,b})\widehat{V}^{*,\nu^*}\big| \le& \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^{*,\nu^*})}{N}}+\frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N}\\
&+\min_{u\in U_s}\big|\widehat{V}^{*,\nu^*}(s)-u\big|\cdot\left(2+\sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right)\\
\big|(P_{s,a,b}-\widehat{P}_{s,a,b})\widehat{V}^{\mu^*,\nu^*}\big| \le& \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^{\mu^*,\nu^*})}{N}}+\frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N}\\
&+\min_{u\in U_s}\big|\widehat{V}^{\mu^*,\nu^*}(s)-u\big|\cdot\left(2+\sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right).
\end{aligned}
$$

*Moreover, recalling that $(\widehat{\mu}_{s,u},\widehat{\nu}_{s,u})$ is the output of the **Planning Oracle** using $\widehat{\mathcal{G}}_{s,u}$, we have*

$$
\begin{aligned}
\big|(P_{s,a,b}-\widehat{P}_{s,a,b})V^{\widehat{\mu},*}\big| \le& \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(V^{\widehat{\mu},*})}{N}}+\frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N}\\
&+\min_{u\in U_s}\big\|V^{\widehat{\mu},*}-V^{\widehat{\mu}_{s,u},*}\big\|_\infty\left(2+\sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right),
\end{aligned}
$$

$$
\begin{aligned}
\big|(P_{s,a,b}-\widehat{P}_{s,a,b})V^{*,\widehat{\nu}}\big| \le& \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(V^{*,\widehat{\nu}})}{N}}+\frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N}\\
&+\min_{u\in U_s}\big\|V^{*,\widehat{\nu}}-V^{*,\widehat{\nu}_{s,u}}\big\|_\infty\left(2+\sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right).
\end{aligned}
$$

Proof First, for all $u \in U_s$ and with probability greater than $1-\delta$, we have

$$
\begin{aligned}
\big|(P_{s,a,b}-\widehat{P}_{s,a,b})\widehat{V}^*\big| &= \big|(P_{s,a,b}-\widehat{P}_{s,a,b})(\widehat{V}^*-\widehat{V}^*_{s,u}+\widehat{V}^*_{s,u})\big| \\
&\le \big|(P_{s,a,b}-\widehat{P}_{s,a,b})(\widehat{V}^*-\widehat{V}^*_{s,u})\big| + \big|(P_{s,a,b}-\widehat{P}_{s,a,b})\widehat{V}^*_{s,u}\big| \qquad (30)\\
&\le 2\cdot\big\|\widehat{V}^*-\widehat{V}^*_{s,u}\big\|_\infty + \big|(P_{s,a,b}-\widehat{P}_{s,a,b})\widehat{V}^*_{s,u}\big| \qquad (31)\\
&\le 2\cdot\big\|\widehat{V}^*-\widehat{V}^*_{s,u}\big\|_\infty + \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^*_{s,u})}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N} \qquad (32)\\
&\le \big\|\widehat{V}^*-\widehat{V}^*_{s,u}\big\|_\infty\left(2+\sqrt{\frac{2\log(4|U_s|/\delta)}{N}}\right) + \sqrt{\frac{2\log(4|U_s|/\delta)\cdot \mathrm{Var}_{P_{s,a,b}}(\widehat{V}^*)}{N}} + \frac{2\log(4|U_s|/\delta)}{3(1-\gamma)N} \qquad (33)
\end{aligned}
$$

where (30)-(31) use triangle inequality, (32) is due to Lemma 13, and (33) uses the facts that $\sqrt{\mathrm{Var}_{P_{s,a,b}}(X+Y)} \le \sqrt{\mathrm{Var}_{P_{s,a,b}}(X)} + \sqrt{\mathrm{Var}_{P_{s,a,b}}(Y)}$, and $\sqrt{\mathrm{Var}_{P_{s,a,b}}(X)} \le \|X\|_\infty$. Moreover, by Lemmas 14 and 15, we obtain that

$$
\big\|\widehat{V}^*-\widehat{V}^*_{s,u}\big\|_\infty = \big\|\widehat{V}^*_{s,\widehat{V}^*(s)}-\widehat{V}^*_{s,u}\big\|_\infty \le \big|\widehat{V}^*(s)-u\big|,
$$

which, combined with (33) and taken minimization over all $u \in U_s$, yields the first inequality. Proofs for the remaining inequalities are analogous, except that for the last two, the norms $\|V^{\widehat{\mu},*}-V^{\widehat{\mu}_{s,u},*}\|_\infty$ and $\|V^{*,\widehat{\nu}}-V^{*,\widehat{\nu}_{s,u}}\|_\infty$ are kept and not further bounded. ■

Next we establish the important result that characterizes the errors $|(P-\widehat{P})\widehat{V}^*|$, $|(P-\widehat{P})\widehat{V}^{\mu^*,*}|$, $|(P-\widehat{P})\widehat{V}^{*,\nu^*}|$, and $|(P-\widehat{P})\widehat{V}^{\mu^*,\nu^*}|$, which could not have been handled without the arguments above, due to the dependence between $\widehat{P}$ and $\widehat{V}^*$ (and also $\widehat{V}^{\mu^*,*}$, $\widehat{V}^{*,\nu^*}$, and $\widehat{V}^{\mu^*,\nu^*}$).

Lemma 17 *For any $\delta \in (0,1]$, with probability greater than $1-\delta$, it holds that*

$$
\begin{aligned}
\big|(P-\widehat{P})\widehat{V}^*\big| &\le \sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)\cdot \mathrm{Var}_P(\widehat{V}^*)}{N}} + \Delta'_{\delta,N}\\
\big|(P-\widehat{P})\widehat{V}^{\mu^*,*}\big| &\le \sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)\cdot \mathrm{Var}_P(\widehat{V}^{\mu^*,*})}{N}} + \Delta'_{\delta,N}\\
\big|(P-\widehat{P})\widehat{V}^{*,\nu^*}\big| &\le \sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)\cdot \mathrm{Var}_P(\widehat{V}^{*,\nu^*})}{N}} + \Delta'_{\delta,N}\\
\big|(P-\widehat{P})\widehat{V}^{\mu^*,\nu^*}\big| &\le \sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)\cdot \mathrm{Var}_P(\widehat{V}^{\mu^*,\nu^*})}{N}} + \Delta'_{\delta,N}
\end{aligned}
$$

where $\Delta'_{\delta,N}$ is defined as

$$
\Delta'_{\delta,N} = \sqrt{\frac{c\log\big(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{N}} + \frac{c\log\big(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{(1-\gamma)N},
$$

and c is some absolute constant.

Proof Let U_s denote a set with evenly spaced elements in the interval $[V^*(s)-\Delta_{\delta/2,N}, V^*(s)+\Delta_{\delta/2,N}]$, with $|U_s|=2/(1-\gamma)^2$, and $\Delta_{\delta,N}$ being defined in Lemma 12. Lemma 12 shows that with probability greater than $1-\delta/2$,

$$\widehat{V}^*(s)\in\big[V^*(s)-\Delta_{\delta/2,N},\quad V^*(s)+\Delta_{\delta/2,N}\big] \tag{34}$$

for all $s\in\mathcal{S}$. Since each subinterval determined by U_s is of length $2\Delta_{\delta/2,N}/(|U_s|-1)$, and $\widehat{V}^*(s)$ will fall into one of them, we know that

$$\min_{u\in U_s}\big|\widehat{V}^*(s)-u\big|\leq\frac{2\Delta_{\delta/2,N}}{|U_s|-1}=\frac{2\gamma}{(|U_s|-1)(1-\gamma)^2}\sqrt{\frac{2\log(4|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}\leq 2\gamma\sqrt{\frac{2\log(4|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}},$$

where we have used the fact that $|U_s|\geq 1/(1-\gamma)^2+1$. We then choose $\delta/2$ to be $\delta/(2|\mathcal{S}||\mathcal{A}||\mathcal{B}|)$ in Lemma 16, so that it holds for all states and joint actions with probability greater than $1-\delta/2$. By substitution and noting that the two events in Lemmas 12 and 16 both fail with probability $\delta/2$, we obtain the first inequality by properly choosing the constant c. Similarly, for the other two inequalities, note that Lemma 12 can be applied to show that $\widehat{V}^{\mu^*,*}(s)$, $\widehat{V}^{*,\nu^*}(s)$, and $\widehat{V}^{\mu^*,\nu^*}(s)$, all lie in the interval in (34) (centered at $V^*(s)$). By similar arguments, the remaining three inequalities can be proved (note that Lemma 16 can be applied to $\widehat{V}^{\mu^*,*}(s)$, $\widehat{V}^{*,\nu^*}(s)$, and $\widehat{V}^{\mu^*,\nu^*}(s)$, as well). ■

Lastly, with a smooth **Planning Oracle**, see Definition 7, we can similarly establish the following error bounds on $|(P-\widehat{P})V^{\widehat{\mu},*}|$ and $|(P-\widehat{P})V^{*,\widehat{\nu}}|$, thanks to Lemma 16.

Lemma 18 *With a smooth **Planning Oracle** that has a smooth constant C (see Definition 7), for any $\delta\in(0,1]$, with probability greater than $1-\delta$, it holds that*

$$\big|(P-\widehat{P})V^{\widehat{\mu},*}\big|\leq\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)\cdot\mathrm{Var}_P(V^{\widehat{\mu},*})}{N}}+\Delta''_{\delta,N}$$

$$\big|(P-\widehat{P})V^{*,\widehat{\nu}}\big|\leq\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)\cdot\mathrm{Var}_P(V^{*,\widehat{\nu}})}{N}}+\Delta''_{\delta,N}$$

where $\Delta''_{\delta,N}$ is defined as

$$\Delta''_{\delta,N}=\sqrt{\frac{c\log\big(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}+\frac{c\log\big(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{(1-\gamma)N},$$

for some absolute constant c.

Proof Following the proof of Lemma 17, let U_s denote a set with evenly spaced elements in the interval $[V^*(s)-\Delta_{\delta/2,N}, V^*(s)+\Delta_{\delta/2,N}]$, with $\Delta_{\delta,N}$ being defined in Lemma 12. By Lemma 12, we know that $\widehat{V}^*(s)$ lies in this interval with probability greater than $1-\delta/2$, for all $s\in\mathcal{S}$. Now we choose $|U_s|=(C+1)/(1-\gamma)^4$, where C is the smooth coefficient in Definition 7. As $\widehat{V}^*(s)$ will fall into one of the subintervals determined by U_s, we have

$$\min_{u\in U_s}\big|\widehat{V}^*(s)-u\big|\leq\frac{2\Delta_{\delta/2,N}}{|U_s|-1}\leq\frac{2\gamma(1-\gamma)^2}{C}\cdot\sqrt{\frac{2\log(4|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}, \tag{35}$$

which also uses the fact $|U_s| \geq C/(1-\gamma)^4 + 1$. Furthermore, by Definition 7 and the proof of Lemma 15, we have

$$\big\|\widehat{\mu} - \widehat{\mu}_{s,u}\big\|_{TV} \leq C \cdot \|\widehat{Q}^* - \widehat{Q}^*_{s,u}\|_\infty \leq C \cdot \big|\widehat{V}^*(s) - u\big|. \tag{36}$$

On the other hand, we have

$$\begin{aligned}
\big\|V^{\widehat{\mu},*} - V^{\widehat{\mu}_{s,u},*}\big\|_\infty &\leq \max_{\vartheta\in\Delta(\mathcal{B})} \big\|\mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}[Q^{\widehat{\mu},*}(\cdot,a,b)] - \mathbb{E}_{a\sim\widehat{\mu}_{s,u}(\cdot\,|\,s),b\sim\vartheta}[Q^{\widehat{\mu}_{s,u},*}(\cdot,a,b)]\big\|_\infty \\
&\leq \max_{\vartheta\in\Delta(\mathcal{B})} \big\|\mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}[Q^{\widehat{\mu},*}(\cdot,a,b)] - \mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}[Q^{\widehat{\mu}_{s,u},*}(\cdot,a,b)]\big\|_\infty \\
&\qquad + \max_{\vartheta\in\Delta(\mathcal{B})} \big\|\mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}[Q^{\widehat{\mu}_{s,u},*}(\cdot,a,b)] - \mathbb{E}_{a\sim\widehat{\mu}_{s,u}(\cdot\,|\,s),b\sim\vartheta}[Q^{\widehat{\mu}_{s,u},*}(\cdot,a,b)]\big\|_\infty \\
&\leq \big\|Q^{\widehat{\mu},*} - Q^{\widehat{\mu}_{s,u},*}\big\|_\infty + \big\|\widehat{\mu} - \widehat{\mu}_{s,u}\big\|_{TV} \cdot \big\|Q^{\widehat{\mu}_{s,u},*}\big\|_\infty \qquad (37) \\
&\leq \gamma\big\|V^{\widehat{\mu},*} - V^{\widehat{\mu}_{s,u},*}\big\|_\infty + \frac{C}{1-\gamma} \cdot \big|\widehat{V}^*(s) - u\big|, \qquad (38)
\end{aligned}$$

where (37) uses Hölder's inequality, and (38) follows by expanding the Q-value functions, using (36), and noticing that $\|Q^{\widehat{\mu}_{s,u},*}\|_\infty \leq 1/(1-\gamma)$. Combining (38) and (35), and taking min over $u \in U_s$, we have

$$\min_{u\in U_s} \big\|V^{\widehat{\mu},*} - V^{\widehat{\mu}_{s,u},*}\big\|_\infty \leq \frac{C}{(1-\gamma)^2} \cdot \min_{u\in U_s} \big|\widehat{V}^*(s) - u\big| \leq 2\gamma \cdot \sqrt{\frac{2\log(4|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}.$$

The rest of the proof follows the arguments of Lemma 17, which combines the last two inequalities in Lemma 16 to obtain the desired bound. Note that the absolute constant here might be different from that in Lemma 17. The proof for the second inequality is analogous. ■

Note that compared to Lemma 17, Lemma 18 has to additionally deal with the interdependence between $\widehat{P}$ and $V^{\widehat{\mu},*}$ (as well as that between $\widehat{P}$ and $V^{*,\widehat{\nu}}$). What can be guaranteed before, in the absorbing MGs, is that the *value* function can be controlled to be close to that in the original MG (see Lemmas 14 and 15, and the proof of Lemma 16). However, in general, it is unclear how much the NE *policy* changes, as well as how much the *best-response value* in the original true MG changes. This calls for some *stability* of the NE policy, and was made possible due to the smoothness of our **Planning Oracle** (see (36)-(38)). Lemma 18 will play an important role in obtaining the near-optimal sample complexity in Theorem 8 (see §4.5).

4.3 Proof of Theorem 5

We are now ready to prove Theorem 5. To this end, we first establish the following lemma.

Lemma 19 *For any policy pair $(\widehat{\mu},\widehat{\nu})$ that satisfies the condition in Theorem 5, there exists some absolute constant c such that*

$$\|Q^{\widehat{\mu},\widehat{\nu}}-\widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty \leq \frac{\gamma}{1-\alpha_{\delta,N}}\left(\sqrt{\frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^3N}}+\frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^2N}\right)$$
$$+\frac{1}{1-\alpha_{\delta,N}}\cdot\frac{\gamma\epsilon_{opt}}{(1-\gamma)}\left(1+\sqrt{\frac{\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{N}}\right)$$
$$\|Q^{*}-\widehat{Q}^{\mu^*,*}\|_\infty \leq \frac{\gamma}{1-\alpha_{\delta,N}}\left(\sqrt{\frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^3N}}+\frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^2N}\right)$$

$$\|Q^{*}-\widehat{Q}^{*,\nu^*}\|_\infty \leq \frac{\gamma}{1-\alpha_{\delta,N}}\left(\sqrt{\frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^3N}}+\frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^2N}\right),$$

where

$$\alpha_{\delta,N}=\frac{\gamma}{1-\gamma}\sqrt{\frac{2\log(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{N}}.$$

Proof Note that

$$\|Q^{\widehat{\mu},\widehat{\nu}}-\widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty=\gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}(P-\widehat{P})\widehat{V}^{\widehat{\mu},\widehat{\nu}}\big\|_\infty \tag{39}$$
$$\leq\gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}(P-\widehat{P})\widehat{V}^{*}\big\|_\infty+\gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}(P-\widehat{P})(\widehat{V}^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^{*})\big\|_\infty \tag{40}$$
$$\leq\gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}\big|(P-\widehat{P})\widehat{V}^{*}\big|\big\|_\infty+\frac{2\gamma\epsilon_{opt}}{1-\gamma}, \tag{41}$$

where (39) is due to Lemma 9; (40) uses triangle inequality; and (41) is due to the non-negativeness of the entries in $(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}$, the sub-optimality of $(\widehat{\mu},\widehat{\nu})$, and Lemma 10. Since the first term in (41) can be bounded using Lemma 17, we have

$$\|Q^{\widehat{\mu},\widehat{\nu}}-\widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty\leq\gamma\sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{N}}\Big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}\sqrt{\mathrm{Var}_P(\widehat{V}^*)}\Big\|_\infty+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}+\frac{2\gamma\epsilon_{opt}}{1-\gamma}$$
$$\leq\gamma\sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{N}}\Big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}\Big(\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},\widehat{\nu}})}+\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^{\widehat{\mu},\widehat{\nu}})}\Big)\Big\|_\infty$$
$$+\gamma\sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{N}}\Big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu}})^{-1}\Big(\sqrt{\mathrm{Var}_P(\widehat{V}^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^*)}\Big)\Big\|_\infty+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}+\frac{2\gamma\epsilon_{opt}}{1-\gamma} \tag{42}$$
$$\leq\gamma\sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{N}}\left(\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\|V^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^{\widehat{\mu},\widehat{\nu}}\|_\infty}{1-\gamma}+\frac{\epsilon_{opt}}{1-\gamma}\right)+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}+\frac{2\gamma\epsilon_{opt}}{1-\gamma} \tag{43}$$

$$\leq \gamma\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\left(\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\|Q^{\widehat{\mu},\widehat{\nu}}-\widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty}{1-\gamma}+\frac{\epsilon_{opt}}{1-\gamma}\right)+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}+\frac{2\gamma\epsilon_{opt}}{1-\gamma} \tag{44}$$

$$= \gamma\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\left(\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\|Q^{\widehat{\mu},\widehat{\nu}}-\widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty}{1-\gamma}\right)+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}$$
$$+\left(2+\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\right)\cdot\frac{\gamma\epsilon_{opt}}{1-\gamma}, \tag{45}$$

where (42) uses the fact that $\sqrt{\mathrm{Var}_P(X+Y)} \leq \sqrt{\mathrm{Var}_P(X)}+\sqrt{\mathrm{Var}_P(Y)}$; (43) is due to Lemma 11, the fact that $\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^{\widehat{\mu},\widehat{\nu}})} \leq \|V^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^{\widehat{\mu},\widehat{\nu}}\|_\infty$, and $\|\widehat{V}^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^*\|_\infty \leq \epsilon_{opt}$; (44) is due to $\|V^{\widehat{\mu},\widehat{\nu}}-\widehat{V}^{\widehat{\mu},\widehat{\nu}}\|_\infty \leq \|Q^{\widehat{\mu},\widehat{\nu}}-\widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty$. Solving for $\|Q^{\widehat{\mu},\widehat{\nu}}-\widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty$ in (45) yields the desired inequality.

For the second inequality, by Lemma 9, we first have

$$\underbrace{\gamma(I-\gamma P^{\mu^*,\nu^*})^{-1}(P-\widehat{P})\widehat{V}^{\mu^*,\nu^*}}_{Q^{\mu^*,\nu^*}-\widehat{Q}^{\mu^*,\nu^*}} \leq Q^*-\widehat{Q}^{\mu^*,*} = Q^{\mu^*,\nu^*}-\widehat{Q}^{\mu^*,*} \leq \underbrace{\gamma(I-\gamma P^{\mu^*,\widehat{\nu(\mu^*)}})^{-1}(P-\widehat{P})\widehat{V}^{\mu^*,*}}_{Q^{\mu^*,\widehat{\nu(\mu^*)}}-\widehat{Q}^{\mu^*,*}}.$$

Thus, we obtain that

$$\left\|Q^*-\widehat{Q}^{\mu^*,*}\right\|_\infty \leq \max\left\{\left\|Q^{\mu^*,\nu^*}-\widehat{Q}^{\mu^*,\nu^*}\right\|_\infty,\ \left\|Q^{\mu^*,\widehat{\nu(\mu^*)}}-\widehat{Q}^{\mu^*,*}\right\|_\infty\right\} \tag{46}$$
$$= \max\left\{\gamma\left\|(I-\gamma P^{\mu^*,\nu^*})^{-1}(P-\widehat{P})\widehat{V}^{\mu^*,\nu^*}\right\|_\infty,\ \gamma\left\|(I-\gamma P^{\mu^*,\widehat{\nu(\mu^*)}})^{-1}(P-\widehat{P})\widehat{V}^{\mu^*,*}\right\|_\infty\right\}.$$

For the first term in the max operator above, by similar arguments from (42)-(45), we have

$$\left\|Q^{\mu^*,\nu^*}-\widehat{Q}^{\mu^*,\nu^*}\right\|_\infty = \gamma\left\|(I-\gamma P^{\mu^*,\nu^*})^{-1}(P-\widehat{P})\widehat{V}^{\mu^*,\nu^*}\right\|_\infty$$
$$\leq \gamma\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\left\|(I-\gamma P^{\mu^*,\nu^*})^{-1}\sqrt{\mathrm{Var}_P(\widehat{V}^{\mu^*,\nu^*})}\right\|_\infty+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma} \tag{47}$$
$$\leq \gamma\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\left\|(I-\gamma P^{\mu^*,\nu^*})^{-1}\sqrt{\mathrm{Var}_P(V^{\mu^*,\nu^*}-\widehat{V}^{\mu^*,\nu^*})}\right\|_\infty$$
$$+\gamma\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\left\|(I-\gamma P^{\mu^*,\nu^*})^{-1}\sqrt{\mathrm{Var}_P(V^{\mu^*,\nu^*})}\right\|_\infty+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma} \tag{48}$$
$$\leq \gamma\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\cdot\frac{\left\|Q^{\mu^*,\nu^*}-\widehat{Q}^{\mu^*,\nu^*}\right\|_\infty}{1-\gamma}$$
$$+\gamma\sqrt{\frac{2\log\left(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\right)}{N}}\cdot\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}, \tag{49}$$

where (47) is due to Lemma 17, (48) uses triangle inequality, and (50) uses Lemma 11. Solving for $\left\|Q^{\mu^*,\nu^*}-\widehat{Q}^{\mu^*,\nu^*}\right\|_\infty$ gives the bound for it.

Similarly, the second term in the max operator in (46) can be bounded by

$$\big\|Q^{\mu^*,\widehat{\nu(\mu^*)}} - \widehat{Q}^{\mu^*,*}\big\|_\infty \leq \gamma\sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{N}} \cdot \frac{\big\|Q^{\mu^*,\widehat{\nu(\mu^*)}} - \widehat{Q}^{\mu^*,*}\big\|_\infty}{1-\gamma} + \gamma\sqrt{\frac{2\log\big(16|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta]\big)}{N}} \cdot \sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\gamma\Delta'_{\delta,N}}{1-\gamma}, \quad (50)$$

which can be solved to obtain a bound for $\big\|Q^{\mu^*,\widehat{\nu(\mu^*)}} - \widehat{Q}^{\mu^*,*}\big\|_\infty$. Combining the two bounds and (46), we prove the second inequality in the lemma. The proof for the third inequality is analogous. ■

With Lemma 19 in hand, we are ready to prove Theorem 5. Note that the condition on N in Theorem 5 makes $\alpha_{\delta,N} < 1/2$. Thus, by (8)-(9) in Lemma 9 with (μ,ν) being replaced by $(\widehat{\mu},\widehat{\nu})$, we have

$$-\|Q^{\widehat{\mu},\widehat{\nu}} - \widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty - \gamma\epsilon_{opt} - \|\widehat{Q}^{\mu^*,*} - Q^*\|_\infty \leq Q^{\widehat{\mu},\widehat{\nu}} - Q^* \leq \|Q^{\widehat{\mu},\widehat{\nu}} - \widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty + \gamma\epsilon_{opt} + \|\widehat{Q}^{*,\nu^*} - Q^*\|_\infty,$$

where we use

$$\|\widehat{Q}^{\widehat{\mu},\widehat{\nu}} - \widehat{Q}^*\|_\infty = \gamma\|P\widehat{V}^{\widehat{\mu},\widehat{\nu}} - P\widehat{V}^*\|_\infty \leq \gamma\|\widehat{V}^{\widehat{\mu},\widehat{\nu}} - \widehat{V}^*\|_\infty \leq \gamma\epsilon_{opt}.$$

Substituting in the bounds of $\|Q^{\widehat{\mu},\widehat{\nu}} - \widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty$, $\|Q^* - \widehat{Q}^{\mu^*,*}\|_\infty$, and $\|Q^* - \widehat{Q}^{*,\nu^*}\|_\infty$ in Lemma 19, we arrive at the final bound for $\|Q^{\widehat{\mu},\widehat{\nu}} - Q^*\|_\infty$:

$$\|Q^{\widehat{\mu},\widehat{\nu}} - Q^*\|_\infty \leq 4\gamma\left(\sqrt{\frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^3N}} + \frac{c\log(c|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^2\delta])}{(1-\gamma)^2N}\right) + \frac{4\gamma\epsilon_{opt}}{1-\gamma} + \gamma\epsilon_{opt}.$$

With a certain choice of c, we have $\|Q^{\widehat{\mu},\widehat{\nu}} - Q^*\|_\infty \leq 2\epsilon/3 + 5\gamma\epsilon_{opt}/(1-\gamma)$.

For the last argument in Theorem 5, by triangle inequality, with the same constant c used above, we have

$$\|\widehat{Q}^{\widehat{\mu},\widehat{\nu}} - Q^*\|_\infty \leq \|Q^{\widehat{\mu},\widehat{\nu}} - Q^*\|_\infty + \|\widehat{Q}^{\widehat{\mu},\widehat{\nu}} - Q^{\widehat{\mu},\widehat{\nu}}\|_\infty \leq \epsilon + \frac{9\gamma\epsilon_{opt}}{1-\gamma},$$

which completes the proof. ■

4.4 Proof of Corollary 6

We now prove Corollary 6, based on Theorem 5. For any state s, we have

$$
\begin{aligned}
V^*(s) - V^{\widetilde{\mu},*}(s) &= \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^{\widetilde{\mu},*}(s,a,b)\big] \\
&= \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] \\
&\quad + \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^{\widetilde{\mu},*}(s,a,b)\big] \\
&\leq \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] + \gamma\|V^* - V^{\widetilde{\mu},*}\|_\infty \qquad (51) \\
&\leq \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^{\widehat{\mu},\widehat{\nu}}(s,a,b)\big] \\
&\quad + \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^{\widehat{\mu},\widehat{\nu}}(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] + \gamma\|V^* - V^{\widetilde{\mu},*}\|_\infty \qquad (52) \\
&\leq 2\big\|Q^* - \widehat{Q}^{\widehat{\mu},\widehat{\nu}}\big\|_\infty + \gamma\|V^* - V^{\widetilde{\mu},*}\|_\infty, \qquad (53)
\end{aligned}
$$

where (51) uses the fact that

$$
\begin{aligned}
&\min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^{\widetilde{\mu},*}(s,a,b)\big] \\
&\quad \leq \max_{\vartheta\in\Delta(\mathcal{B})} \Big|\mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^*(s,a,b)\big] - \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[Q^{\widetilde{\mu},*}(s,a,b)\big]\Big| \leq \gamma\|V^* - V^{\widetilde{\mu},*}\|_\infty,
\end{aligned}
$$

and (52) is due to the fact that

$$
-\min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\mu^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^{\widehat{\mu},\widehat{\nu}}(s,a,b)\big] + \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widetilde{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^{\widehat{\mu},\widehat{\nu}}(s,a,b)\big] \geq 0,
$$

by definition of $\widetilde{\mu}$. Hence, (53), together with Theorem 5, implies that

$$
V^* - V^{\widetilde{\mu},*} \leq \frac{2\big\|Q^* - \widehat{Q}^{\widehat{\mu},\widehat{\nu}}\big\|_\infty}{1-\gamma} = \widetilde{\epsilon}. \qquad (54)
$$

By similar arguments, we have

$$
V^{*,\widetilde{\nu}} - V^* \leq \frac{2\big\|Q^* - \widehat{Q}^{\widehat{\mu},\widehat{\nu}}\big\|_\infty}{1-\gamma} = \widetilde{\epsilon}. \qquad (55)
$$

Combining (54) and (55) yields

$$
V^{\widetilde{\mu},\widetilde{\nu}} - V^{\widetilde{\mu},*} \leq V^{*,\widetilde{\nu}} - V^{\widetilde{\mu},*} \leq 2\widetilde{\epsilon}, \qquad V^{*,\widetilde{\nu}} - V^{\widetilde{\mu},\widetilde{\nu}} \leq V^{*,\widetilde{\nu}} - V^{\widetilde{\mu},*} \leq 2\widetilde{\epsilon},
$$

which completes the proof. ■

4.5 Proof of Theorem 8

We now prove the second main result, Theorem 8. First, following the proof of Corollary 6, it suffices to prove that $V^* - V^{\widehat{\mu},*} \leq \widetilde{\epsilon},\ \ V^{*,\widehat{\nu}} - V^* \leq \widetilde{\epsilon}$, since they together imply that $(\widehat{\mu},\widehat{\nu})$ is a $2\widetilde{\epsilon}$-Nash equilibrium. The following analysis is devoted to proving this argument.

The idea is similar to that presented in §4.3, i.e., we use the component-wise error decompositions in Lemma 9, but use (10)-(11) instead. In particular, letting $\mu = \widehat{\mu}$ and $\nu = \widehat{\nu}$, we have

$$V^{\widehat{\mu},*} - V^* \geq -\|Q^{\widehat{\mu},*} - \widehat{Q}^{\widehat{\mu},*}\|_\infty - \epsilon_{opt} - \|\widehat{Q}^{\mu^*,*} - Q^*\|_\infty \tag{56}$$

$$V^{*,\widehat{\nu}} - V^* \leq \|Q^{*,\widehat{\nu}} - \widehat{Q}^{*,\widehat{\nu}}\|_\infty + \epsilon_{opt} + \|\widehat{Q}^{*,\nu^*} - Q^*\|_\infty. \tag{57}$$

Note that the bounds for $\|\widehat{Q}^{\mu^*,*} - Q^*\|_\infty$ and $\|\widehat{Q}^{*,\nu^*} - Q^*\|_\infty$ have already been established in Lemma 19 (without dependence on ϵ_{opt} and the **Planning Oracle**). It now suffices to bound $\|Q^{\widehat{\mu},*} - \widehat{Q}^{\widehat{\mu},*}\|_\infty$ and $\|Q^{*,\widehat{\nu}} - \widehat{Q}^{*,\widehat{\nu}}\|_\infty$. For the former term, by Lemma 9, we first have

$$\underbrace{\gamma(I-\gamma\widehat{P}^{\widehat{\mu},\nu(\widehat{\mu})})^{-1}(P-\widehat{P})V^{\widehat{\mu},\nu(\widehat{\mu})}}_{Q^{\widehat{\mu},*}-\widehat{Q}^{\widehat{\mu},\nu(\widehat{\mu})}} \leq Q^{\widehat{\mu},*} - \widehat{Q}^{\widehat{\mu},*} \leq \underbrace{\gamma(I-\gamma P^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}})^{-1}(P-\widehat{P})\widehat{V}^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}}}_{Q^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}}-\widehat{Q}^{\widehat{\mu},*}}.$$

Thus, we know that

$$\begin{aligned}
&\big\|Q^{\widehat{\mu},*} - \widehat{Q}^{\widehat{\mu},*}\big\|_\infty \qquad (58)\\
&\leq \max\Big\{\gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}})^{-1}(P-\widehat{P})\widehat{V}^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}}\big\|_\infty,\ \ \gamma\big\|(I-\gamma\widehat{P}^{\widehat{\mu},\nu(\widehat{\mu})})^{-1}(P-\widehat{P})V^{\widehat{\mu},\nu(\widehat{\mu})}\big\|_\infty\Big\}.
\end{aligned}$$

The first term in the max operator, where the policies in the pair $(\widehat{\mu},\widehat{\nu(\widehat{\mu})})$ are both obtained from the empirical model $\widehat{\mathcal{G}}$, can be bounded similarly as that for $\|Q^{\widehat{\mu},\widehat{\nu}} - \widehat{Q}^{\widehat{\mu},\widehat{\nu}}\|_\infty$ in Lemma 19. Specifically, following (39)-(41), we have

$$\begin{aligned}
&\gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}})^{-1}(P-\widehat{P})\widehat{V}^{\widehat{\mu},*}\big\|_\infty\\
&\quad\leq \gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}})^{-1}(P-\widehat{P})\widehat{V}^*\big\|_\infty + \gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}})^{-1}(P-\widehat{P})(\widehat{V}^{\widehat{\mu},*}-\widehat{V}^*)\big\|_\infty \qquad (59)\\
&\quad\leq \gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}})^{-1}\big|(P-\widehat{P})\widehat{V}^*\big|\big\|_\infty + \frac{2\gamma\epsilon_{opt}}{1-\gamma}, \qquad (60)
\end{aligned}$$

where (59) uses triangle inequality, and (60) is due to the optimization error of $\widehat{\mu}$. Then, to bound $\gamma\big\|(I-\gamma P^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}})^{-1}\big|(P-\widehat{P})\widehat{V}^*\big|\big\|_\infty$, the rest of the proof is analogous to the derivations in (42)-(45), by replacing $\widehat{\nu}$ therein by $\widehat{\nu(\widehat{\mu})}$, and bound $\|\widehat{V}^{\widehat{\mu},*} - \widehat{V}^*\|_\infty$ by ϵ_{opt}. Solving for $\|Q^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}} - \widehat{Q}^{\widehat{\mu},*}\|_\infty$ yields the desired bound for the first term in the max in (58), namely, there exists some constant c such that with probability greater than $1-\delta$,

$$\begin{aligned}
&\big\|Q^{\widehat{\mu},\widehat{\nu(\widehat{\mu})}} - \widehat{Q}^{\widehat{\mu},*}\big\|_\infty\\
&\leq \frac{\gamma}{1-\alpha'_{\delta,N}}\left(\sqrt{\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^3N}} + \frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^2N}\right)\\
&\quad + \frac{1}{1-\alpha'_{\delta,N}}\cdot\frac{\gamma\epsilon_{opt}}{(1-\gamma)}\left(1+\sqrt{\frac{\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{N}}\right), \qquad (61)
\end{aligned}$$

where $\alpha'_{\delta,N}$ is defined as

$$\alpha'_{\delta,N} = \frac{\gamma}{1-\gamma}\sqrt{\frac{2\log(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{N}}.$$

For the second term in the max in (58), note that $\widehat{\mu}$ is obtained from $\widehat{\mathcal{G}}$, while $\nu(\widehat{\mu})$ is obtained from the true model $\mathcal{G}$. Note that this mismatch is one key difference from the single-agent setting (Agarwal et al., 2019a) and the above proof for the first term. By Lemma 18, it holds that

$$\begin{aligned}
&\gamma\big\|(I-\gamma\widehat{P}^{\widehat{\mu},\nu(\widehat{\mu})})^{-1}\big|(P-\widehat{P})V^{\widehat{\mu},*}\big|\big\|_\infty\\
&\leq \gamma\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}\Big\|(I-\gamma\widehat{P}^{\widehat{\mu},\nu(\widehat{\mu})})^{-1}\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},*})}\Big\|_\infty+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}\\
&\leq \gamma\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}\cdot\bigg[\Big\|(I-\gamma\widehat{P}^{\widehat{\mu},\nu(\widehat{\mu})})^{-1}\Big(\sqrt{\mathrm{Var}_{\widehat{P}}(\widehat{V}^{\widehat{\mu},\nu(\widehat{\mu})})} \qquad (62)\\
&\quad+\sqrt{\mathrm{Var}_{\widehat{P}}(V^{\widehat{\mu},*}-\widehat{V}^{\widehat{\mu},\nu(\widehat{\mu})})}\Big)\Big\|_\infty+\Big\|(I-\gamma\widehat{P}^{\widehat{\mu},\nu(\widehat{\mu})})^{-1}\Big|\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},*})}-\sqrt{\mathrm{Var}_{\widehat{P}}(V^{\widehat{\mu},*})}\Big|\Big\|_\infty\bigg]+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}\\
&\leq \gamma\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}\bigg(\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\|Q^{\widehat{\mu},*}-\widehat{Q}^{\widehat{\mu},\nu(\widehat{\mu})}\|_\infty}{1-\gamma}\bigg)+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma} \qquad (63)\\
&\quad+\frac{\gamma}{1-\gamma}\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}\Big\|\Big|\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},*})}-\sqrt{\mathrm{Var}_{\widehat{P}}(V^{\widehat{\mu},*})}\Big|\Big\|_\infty,
\end{aligned}$$

where (62) uses the norm-like triangle-inequality property of $\sqrt{\mathrm{Var}_P(V)}$ and triangle inequality, (63) is due to Lemma 11, and the facts that $\sqrt{\mathrm{Var}_P(X)}\leq\|X\|_\infty$, $\|V^{\widehat{\mu},*}-\widehat{V}^{\widehat{\mu},\nu(\widehat{\mu})}\|_\infty\leq\|Q^{\widehat{\mu},*}-\widehat{Q}^{\widehat{\mu},\nu(\widehat{\mu})}\|_\infty$, and Lemma 10. Moreover, notice that

$$\begin{aligned}
&\Big\|\Big|\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},*})}-\sqrt{\mathrm{Var}_{\widehat{P}}(V^{\widehat{\mu},*})}\Big|\Big\|_\infty\\
&\quad\leq\Big\|\Big|\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},*})}-\sqrt{\mathrm{Var}_P(V^*)}\Big|\Big\|_\infty+\Big\|\Big|\sqrt{\mathrm{Var}_{\widehat{P}}(V^{\widehat{\mu},*})}-\sqrt{\mathrm{Var}_{\widehat{P}}(V^*)}\Big|\Big\|_\infty\\
&\qquad+\Big\|\Big|\sqrt{\mathrm{Var}_P(V^*)}-\sqrt{\mathrm{Var}_{\widehat{P}}(V^*)}\Big|\Big\|_\infty \qquad (64)\\
&\quad\leq\Big\|\sqrt{\mathrm{Var}_P(V^{\widehat{\mu},*}-V^*)}\Big\|_\infty+\Big\|\sqrt{\mathrm{Var}_{\widehat{P}}(V^{\widehat{\mu},*}-V^*)}\Big\|_\infty+\Big\|\sqrt{\Big|\mathrm{Var}_P(V^*)-\mathrm{Var}_{\widehat{P}}(V^*)\Big|}\Big\|_\infty \qquad (65)\\
&\quad\leq 2\big\|V^{\widehat{\mu},*}-V^*\big\|_\infty+\sqrt{\Big\|\mathrm{Var}_P(V^*)-\mathrm{Var}_{\widehat{P}}(V^*)\Big\|_\infty}, \qquad (66)
\end{aligned}$$

where (64) uses triangle inequality, (65) uses the norm-like triangle inequality of $\sqrt{\mathrm{Var}_P(V)}$ and $\sqrt{\mathrm{Var}_{\widehat{P}}(V)}$, and the fact $|\sqrt{X}-\sqrt{Y}|\leq\sqrt{|X-Y|}$ for $X,Y\geq 0$, and (66) uses $\sqrt{\mathrm{Var}_P(X)}\leq\|X\|_\infty$ and the definition of $\|\cdot\|_\infty$. In addition, we know that with probability at least $1-\delta$,

$$\Big\|\operatorname{Var}_P(V^*)-\operatorname{Var}_{\widehat{P}}(V^*)\Big\|_\infty=\Big\|(P-\widehat{P})(V^*)^2-\Big((PV^*)^2-(\widehat{P}V^*)^2\Big)\Big\|_\infty$$
$$\leq\Big\|(P-\widehat{P})(V^*)^2\Big\|_\infty+\Big\|(PV^*)^2-(\widehat{P}V^*)^2\Big\|_\infty$$
$$\leq\frac{1}{(1-\gamma)^2}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}+\frac{2}{1-\gamma}\|(P-\widehat{P})V^*\|_\infty\leq\frac{3}{(1-\gamma)^2}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}},\tag{67}$$

due to Hoeffding bound and $\|V^*\|_\infty\leq 1/(1-\gamma)$. Combining (63), (66), and (67) yields

$$\Big\|Q^{\widehat{\mu},*}-\widehat{Q}^{\widehat{\mu},\nu(\widehat{\mu})}\Big\|_\infty$$
$$\leq\gamma\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}\bigg(\sqrt{\frac{2}{(1-\gamma)^3}}+\frac{\|Q^{\widehat{\mu},*}-\widehat{Q}^{\widehat{\mu},\nu(\widehat{\mu})}\|_\infty}{1-\gamma}\bigg)+\frac{\gamma\Delta'_{\delta,N}}{1-\gamma}$$
$$+\frac{\gamma}{1-\gamma}\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}\Big(2\big\|V^{\widehat{\mu},*}-V^*\big\|_\infty$$
$$+\sqrt{\frac{3}{(1-\gamma)^2}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}}\Big).$$

Solving for $\|Q^{\widehat{\mu},*}-\widehat{Q}^{\widehat{\mu},\nu(\widehat{\mu})}\|_\infty$ further leads to

$$\|Q^{\widehat{\mu},*}-\widehat{Q}^{\widehat{\mu},\nu(\widehat{\mu})}\|_\infty$$
$$\leq\frac{\gamma}{1-\alpha'_{\delta,N}}\left(\sqrt{\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^3N}}+\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^2N}\right)$$
$$+\frac{1}{1-\alpha'_{\delta,N}}\cdot\frac{\gamma}{1-\gamma}\sqrt{\frac{2\log\big(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\big)}{N}}\Big(2\big\|V^{\widehat{\mu},*}-V^*\big\|_\infty$$
$$+\frac{1}{1-\gamma}\sqrt[4]{\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}\Big),\tag{68}$$

for some absolute constant c.

Now we substitute (61) and (68) into (58), to complete the bound in (56). If the first term in the max in (58) is larger, and noticing that the choice of N in the theorem can make $\alpha'_{\delta,N}<1/5$, (56), (58), (61), and Lemma 19 together lead to

$$V^*-V^{\widehat{\mu},*}\leq\frac{5\gamma}{2}\left(\sqrt{\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^3N}}+\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^2N}\right)$$
$$+\frac{5\gamma\epsilon_{opt}}{2(1-\gamma)}+\epsilon_{opt},\tag{69}$$

with some absolute constant c, where we have replaced the term $\log(1/(1-\gamma)^2)$ in the bounds for $\|Q^*-\widehat{Q}^{\mu^*,*}\|_\infty$ and $\|Q^*-\widehat{Q}^{*,\nu^*}\|_\infty$ in Lemma 19 (including that in the definition of $\alpha_{\delta,N}$) by $\log((C+1)/(1-\gamma)^4)$, a larger number. If the second term in the max in (58) is larger, (56), (58), (68), and Lemma 19 together yield

$$V^* - V^{\widehat{\mu},*} \le \frac{5\gamma}{2}\left(\sqrt{\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^3 N}} + \frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^2 N}\right)$$
$$+ \frac{5}{4}\cdot\frac{\gamma}{1-\gamma}\sqrt{\frac{2\log\left(8(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta]\right)}{N}}\Big(2\big\|V^{\widehat{\mu},*} - V^*\big\|_\infty$$
$$+ \frac{1}{1-\gamma}\sqrt[4]{\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}}\Big) + \epsilon_{opt},$$

where we have used the fact that $\alpha'_{\delta,N} < 1/5$. Taking infinity norm on both sides and solving for $\|V^{\widehat{\mu},*} - V^*\|_\infty$, we have

$$V^* - V^{\widehat{\mu},*} \le \big\|V^{\widehat{\mu},*} - V^*\big\|_\infty \le 5\gamma\left(\sqrt{\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^3 N}} + \right. \tag{70}$$
$$\left.\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/[(1-\gamma)^4\delta])}{(1-\gamma)^2 N}\right) + \frac{5\gamma}{2(1-\gamma)^2}\Big(\frac{c\log(c(C+1)|\mathcal{S}||\mathcal{A}||\mathcal{B}|/\delta)}{N}\Big)^{3/4} + 2\epsilon_{opt},$$

with some absolute constant c (which can be different from that in (69)). Using the choice of N in the theorem, and combining (69) and (70), we finally have $V^* - V^{\widehat{\mu},*} \le \epsilon + 4\epsilon_{opt}/(1-\gamma)$. Note that on the right-hand side of (70), the N that makes the third term to be $\mathcal{O}(\epsilon)$ is $\widetilde{\mathcal{O}}(1/[(1-\gamma)^{8/3}\epsilon^{4/3}])$, which is dominated by $\widetilde{\mathcal{O}}(1/[(1-\gamma)^3\epsilon^2])$ when $\epsilon \in (0, 1/(1-\gamma)^{1/2}]$. In addition, to make $\alpha'_{\delta,N} < 1/5$, N should be larger than $\widetilde{\mathcal{O}}(1/(1-\gamma)^2)$, which is consistent with both the first and third terms on the right-hand side of (70) to be $\widetilde{\mathcal{O}}(1/(1-\gamma)^{1/2})$, determining the allowed range of ϵ to be $(0, 1/(1-\gamma)^{1/2}]$. This proves the first bound in the theorem.

The proof for completing the bound in (57) is analogous: using Lemmas 18 and 9 to bound $\|Q^{*,\widehat{\nu}} - \widehat{Q}^{*,\widehat{\nu}}\|_\infty$, which is then substituted into (57). This completes the proof. ■

5. Concluding Remarks

In this paper, we have established the first (near-)minimax optimal sample complexity for model-based MARL, when a generative model is available. Our setting was focused on the basic model in MARL — infinite-horizon discounted two-player zero-sum Markov games (Littman, 1994). By noticing that reward is not used in the sampling process of this model-based approach, we have separated the reward-aware and reward-agnostic cases, and established sample complexity lower bounds correspondingly, a unique separation in the multi-agent context. We have then shown that this simple model-based approach is near-minimax optimal in the reward-aware case, with only a gap in the dependence on $|\mathcal{A}|, |\mathcal{B}|$; and is indeed minimax-optimal in the reward-agnostic case. This separation and the (near-)optimal results have not only justified the sample-efficiency of this simple approach, but also reflected both its power (easily handling multiple reward functions known in hindsight), and its limitation (less adaptive and can hardly achieve the optimal $\widetilde{\mathcal{O}}(|\mathcal{A}|+|\mathcal{B}|)$). We believe that our results may shed light on the choice of model-free and model-based approaches in various MARL scenarios in practice.

Our results naturally open up the following interesting future directions. First, besides the turn-based setting in Sidford et al. (2020) and the episodic setting in the concurrent work Bai et al. (2020), the minimax-optimal sample complexity in *all* parameters for *model-free* algorithms is still open. As discussed in §3, in the reward-aware case, the $\widetilde{\Omega}(|\mathcal{A}|+|\mathcal{B}|)$ lower bound may only be attainable by model-free ones. It would be interesting to compare the results with our model-based ones, in both reward-aware and reward-agnostic cases, to better understand their pros and cons in various MARL settings. It would also be interesting to explore the (near-)optimal sample complexity or regret of model-based approaches in other MARL scenarios, such as when no generative model is available, episodic and average-reward settings, general-sum Markov games, and the setting with function approximation.

Acknowledgments

The research of K.Z. and T.B. was supported in part by the US Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196, and in part by the Office of Naval Research (ONR) MURI Grant N00014-16-1-2710. The research of S.K. was supported by the funding from the ONR award N00014-18-1-2247, and NSF Awards CCF-1703574 and CCF-1740551. We would also like to thank all the anonymous reviewers for their valuable feedback that helped improve our paper.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. *arXiv preprint arXiv:1906.03804*, 2019a.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019b.

Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning-Volume 70*, pages 263–272, 2017.

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

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

Ronen I Brafman and Moshe Tennenholtz. R-MAX-A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multi-agent reinforcement learning. *IEEE Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews, 38 (2), 2008*, 2008.

Ashish Cherukuri, Enrique Mallada, Steven Low, and Jorge Cortés. The role of convexity in saddle-point dynamics: Lyapunov function and robustness. *IEEE Transactions on Automatic Control*, 63(8):2449–2464, 2017.

Yinlam Chow, Ofir Nachum, and Mohammad Ghavamzadeh. Path consistency learning in Tsallis entropy regularized MDPs. In *International Conference on Machine Learning*, pages 979–988, 2018.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

George Bernard Dantzig. *Linear programming and extensions*, volume 48. Princeton University Press, 1998.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

Jianqing Fan, Zhuoran Yang, Yuchen Xie, and Zhaoran Wang. A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*, 2019.

John Fearnley and Rahul Savani. Finding approximate Nash equilibria of bimatrix games via payoff queries. *ACM Transactions on Economics and Computation*, 4(4):1–19, 2016.

John Fearnley, Martin Gairing, Paul W. Goldberg, and Rahul Savani. Learning equilibria of games via payoff queries. *Journal of Machine Learning Research*, 16(39):1305–1344, 2015.

Fei Feng, Wotao Yin, and Lin F Yang. Does knowledge transfer always help to learn a better policy? *arXiv preprint arXiv:1912.02986*, 2019.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169, 2019.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6): 359–483, 2015.

Amy Greenwald, Keith Hall, and Roberto Serrano. Correlated Q-learning. In *International Conference on Machine Learning*, volume 20, page 242, 2003.

Jean-Bastien Grill, Omar Darwiche Domingues, Pierre Ménard, Rémi Munos, and Michal Valko. Planning in entropy-regularized Markov decision processes and games. In *Advances in Neural Information Processing Systems*, pages 12383–12392, 2019.

Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

Junling Hu and Michael P Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4(Nov):1039–1069, 2003.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

MJM Jansen. Regularity and stability of equilibrium points of bimatrix games. *Mathematics of Operations Research*, 6(4):530–550, 1981.

Zeyu Jia, Lin F Yang, and Mengdi Wang. Feature-based Q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – A simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

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

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Annual ACM Symposium on Theory of Computing*, pages 302–311, 1984.

Michael J Kearns and Satinder P Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems*, pages 996–1002, 1999.

Steven G Krantz and Harold R Parks. *The Implicit Function Theorem: History, Theory, and Applications.* Springer Science & Business Media, 2012.

Tor Lattimore and Marcus Hutter. PAC bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. In *International Conference on Learning Representations*, 2021.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *arXiv preprint arXiv:2005.12900*, 2020.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *International Conference on Artificial Intelligence and Statistics*, pages 907–915, 2019.

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

Michael L Littman. Friend-or-foe Q-learning in general-sum games. In *International Conference on Machine Learning*, volume 1, pages 322–328, 2001.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

R Duncan Luce and Howard Raiffa. *Games and Decisions: Introduction and Critical Survey*. Courier Corporation, 1989.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

Panayotis Mertikopoulos and William H Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Remi Munos and Andrew W Moore. Variable resolution discretization for high-accuracy solutions of optimal control problems. 1999.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Thanh Thi Nguyen and Vijay Janapa Reddi. Deep reinforcement learning for cyber security. *arXiv preprint arXiv:1906.05799*, 2019.

Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, 2020.

OpenAI. Openai five. `https://blog.openai.com/openai-five/`, 2018.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

Martin J Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT press, 1994.

Stephen David Patek. *Stochastic and Shortest Path Games: Theory and Algorithms*. PhD thesis, Massachusetts Institute of Technology, 1997.

Julien Pérolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pages 1321–1329, 2015.

Julien Pérolat, Bilal Piot, Bruno Scherrer, and Olivier Pietquin. On the use of non-stationary strategies for solving two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 893–901, 2016a.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning Nash equilibrium for general-sum Markov games from batch data. *arXiv preprint arXiv:1606.08718*, 2016b.

Julien Pérolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Muhammed O Sayin, Kaiqing Zhang, and Asuman Ozdaglar. Fictitious play in Markov games with single controller. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 919–936, 2022.

Devavrat Shah, Varun Somani, Qiaomin Xie, and Zhi Xu. On reinforcement learning for turn-based zero-sum Markov games. *arXiv preprint arXiv:2002.10620*, 2020.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

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

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196, 2018.

Aaron Sidford, Mengdi Wang, Lin F Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

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

Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in Neural Information Processing Systems*, pages 3422–3435, 2018.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M. Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, Timo Ewalds, Dan Horgan, Manuel Kroiss, Ivo Danihelka, John Agapiou, Junhyuk Oh, Valentin Dalibard, David Choi, Laurent Sifre, Yury Sulsky, Sasha Vezhnevets, James Molloy, Trevor Cai, David Budden, Tom Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Toby Pohlen, Yuhuai Wu, Dani Yogatama, Julia Cohen, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Chris Apps, Koray Kavukcuoglu, Demis Hassabis, and David Silver. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. `https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/`, 2019.

John Von Neumann, Oskar Morgenstern, and Harold William Kuhn. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

Mengdi Wang. Randomized linear programming solves the discounted Markov decision problem in nearly-linear running time. *arXiv preprint arXiv:1704.01869*, 2017.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5867–5876, 2018a.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1812.02783*, 2018b.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11598–11610, 2019a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021b.

Xiangyuan Zhang, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Non-cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 9482–9493, 2019b.

Appendix A. Lower Bounds

Now we discuss lower bounds of the sample complexity given in §3.1.

A.1 Reward-Aware Case

Proof of Lemma 3. The proof follows by recalling the hard cases of MDPs considered in Azar et al. (2013) or Feng et al. (2019), and replacing each action a therein by a joint-action (a, b). Without loss of generality, suppose $|\mathcal{A}| \geq |\mathcal{B}|$. Then, we design a Markov game such that agent 2 has no effect on the reward or the transition. Thus, finding an NE is now the same as agent 1 finding the optimal value/policy. By the arguments in Azar et al. (2013); Feng et al. (2019), the sample complexity is at least $\widetilde{\Omega}\big(|\mathcal{S}| \cdot \max\{|\mathcal{A}|, |\mathcal{B}|\} \cdot (1-\gamma)^{-3}\epsilon^{-2}\big)$, where $\widetilde{\Omega}$ suppresses some log factors of $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{B}|$ and $1/\delta$. Noticing that $\max\{|\mathcal{A}|, |\mathcal{B}|\} = (|\mathcal{A}| + |\mathcal{B}| + \big||\mathcal{A}| - |\mathcal{B}|\big|)/2$, we obtain the lower bound. ■

Challenge in Obtaining $\widetilde{\Omega}(|\mathcal{A}||\mathcal{B}|)$. Note that the proof of a $\widetilde{\Omega}\big(|\mathcal{S}|(|\mathcal{A}| + |\mathcal{B}|) \cdot (1-\gamma)^{-3}\epsilon^{-2}\big)$ lower bound is a straightforward adaptation from the single-agent result. The lower bound can also be obtained in several other ways (via a treatment of turn-based Markov games, or the attempts to be introduced next). Nevertheless, these attempts can hardly lead to a lower bound of $\widetilde{\Omega}(|\mathcal{A}||\mathcal{B}|)$, in this reward-aware case. We highlight the challenges as follows.

The core proof idea of Azar et al. (2013); Feng et al. (2019) for the single-agent setting lower bound is to create a class of $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ number of MDPs, which are hard to distinguish from each other. When the reward function is given (i.e., in the reward-aware setting), one can only change the transition model to obtain different hard MDPs. Hence, in Azar et al. (2013), their approach is to first create a null hypothesis, in which the optimal Q-value and ϵ-optimal actions at every state are fixed. Then in each of the $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ alternative hypothesis, they change the transition probability of a distinct state-action pair (s, a) in the null case to make the Q-value slightly differ from the null-setting and such that a is an ϵ-optimal action at state s. They construct the hard instance cleverly such that if an algorithm correctly outputs the optimal Q-value (or optimal policy) in an alternative hypothesis with high probability, then it must have sampled $\widetilde{\Omega}((1-\gamma)^{-3}\epsilon^{-2})$ samples at the corresponding (s,a) pair in the null hypothesis. As this holds for all $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ alternative hypotheses, we obtain an $\widetilde{\Omega}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\epsilon^{-2})$ sample lower bound.

In the game setting, however, the above idea requires to change the Nash equilibrium (say, a unique pure strategy) to a different state-action-action tuple at any state while only make changes to the probability transition of the corresponding state-joint-action tuple. Nevertheless, this is challenging to achieve in general, as the NE value of zero-sum matrix games is *not sensitive* to the small number of element changes in the payoff matrices. This can be evidenced either by the stability of the NE in this case against the payoff perturbation (Jansen, 1981), or by the sensitivity analysis of the equivalent linear program of the game (Luce and Raiffa, 1989) against the problem data (Dantzig, 1998). Indeed, one can verify that only changing $\mathcal{O}(1)$ elements in the transition probability matrix, and thus changing $\mathcal{O}(1)$ elements in the Q-value table at each state, by a small amount, can hardly change the NE value/policy too much. Some order of $\mathcal{O}(|\mathcal{A}|)$ (or $\mathcal{O}(|\mathcal{B}|)$) number of changes may suffice, but will eventually yields $\mathcal{O}(|\mathcal{B}|)$ (or $\mathcal{O}(|\mathcal{A}|)$) hard alternative cases, leading to the same $\widetilde{\Omega}(|\mathcal{A}|+|\mathcal{B}|)$ result as Lemma 3. In other words, one can hardly obtain the sufficient number of required hard cases ($\widetilde{\Omega}(|\mathcal{A}||\mathcal{B}|)$ in total) by changing only $\mathcal{O}(1)$ elements in the transition probability matrix of each alternative hypothesis case.

On the other hand, interestingly, we note that there are some results on the *payoff query complexity*, i.e., the number of queries for the elements in the payoff matrix, for finding the NE (Fearnley et al., 2015; Fearnley and Savani, 2016). It is possible to use $\mathcal{O}(k\log(k)/\epsilon^2)$ queries to find the ϵ-NE in zero-sum matrix games when $|\mathcal{A}|=|\mathcal{B}|$, where $k=|\mathcal{A}|=|\mathcal{B}|$ (Fearnley and Savani, 2016). Note that the lower bound given in Fearnley and Savani (2016), though being $\Omega(k^2)$, requires the accuracy $\epsilon\leq 1/k$ to be small, which cannot be used in our previous analysis with a dimension-free choice of ϵ. From a different angle, these results imply that it may indeed be unnecessary to accurately estimate *all* elements in the matrix, in order to obtain an approximate Nash equilibrium.

In light of these observations, we have conjectured that with reward knowledge, the lower bound of $\widetilde{\Omega}(|\mathcal{A}|+|\mathcal{B}|)$ is indeed unimprovable, which might be matched by some other (possibly *model-free*) MARL algorithms, as general model-based approaches inherently require $\widetilde{\Omega}(|\mathcal{A}||\mathcal{B}|)$ for transition model estimation. Such a $\widetilde{\Omega}(|\mathcal{A}|+|\mathcal{B}|)$ lower bound on regret has been provided recently in Bai and Jin (2020), though in a different setting. More interestingly, though not entirely comparable to us, in the concurrent work Bai et al. (2020), the $\widetilde{\mathcal{O}}(|\mathcal{A}|+|\mathcal{B}|)$ complexity is indeed shown to be attainable by a model-free Nash-V learning algorithm in the episodic setting, with the reward information guiding the online update.

A.2 Reward-Agnostic Case

Now we establish the lower bound for the reward-agnostic case, i.e., the proof of Theorem 4. The idea to construct hard cases is similar to that discussed in §A.1, which is motivated by Azar et al. (2013); Feng et al. (2019), but with additional flexibility to design the reward function that is unknown in the sampling stage. Our hard cases apply to both finding the ϵ-NE policy pair and finding the ϵ-approximate NE value. For the sake of presentation, we focus on proving the lower bound for the ϵ-NE policy. Let us first formally define the notion of a correct algorithm in terms of learning an ϵ-NE policy in this reward-agnostic case.

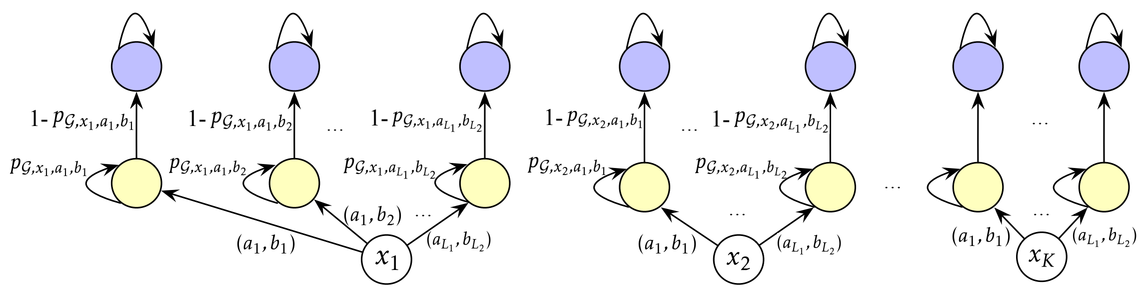

Figure 1: The class of zero-sum Markov games $\mathbb{G}$ considered in the proof of Theorem 4. Circles denote the states and arrows denote the transitions. White, yellow, and blue circles denote the three disjoint subsets of states $\mathcal{X}$, $\mathcal{Y}_1$, and $\mathcal{Y}_2$, respectively.

Definition 20 *((ϵ, δ)-correct reward-agnostic algorithm) We say that an RL algorithm $\mathfrak{A}$ is (ϵ, δ)-correct in the reward-agnostic case, if for any unknown MG $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, r, P, \gamma)$, $\mathfrak{A}$ first calls a generative model on $(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, \gamma)$, and is then fed with the reward r, and outputs an ϵ-NE policy (μ, ν) with probability at least $1 - \delta$.*

Note that r is only revealed to $\mathfrak{A}$ after the sampling, and such an $\mathfrak{A}$ should be able to output an (ϵ, δ)-correct NE policy for any single r in the underlying model. Thus, for M reward functions defined over the same $(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, \gamma)$, using a union bound argument, the ϵ-NE policy corresponding to all M reward functions can be obtained simultaneously with probability greater than $1 - M\delta$ (of course with a small enough δ). To prove the theorem, we will construct a class of Markov game models. We show that if algorithm $\mathfrak{A}$ only draws samples much fewer than the lower bound, there exists an MG $\mathcal{G}$ such that $\mathfrak{A}$ cannot be an (ϵ, δ)-correct reward-agnostic algorithm for. Compared to the reward-aware case, we now allow more freedom to construct hard instances, by not only perturbing the transition matrix, but also choosing the reward function judiciously. This would eventually allow us to obtain $\Theta(|\mathcal{A}||\mathcal{B}|)$ hard cases, combating the insensitivity of NE to the perturbation of the payoff matrices (c.f. discussion in §A.1).

Construction of the Hard Case. We define a family of MGs $\mathbb{G}$. See illustrations in Figure 1. The state space $\mathcal{S}$ consists of three disjoint subsets $\mathcal{X}$, $\mathcal{Y}_1$, and $\mathcal{Y}_2$. The set $\mathcal{X}$ includes K states $\{x_1, x_2, \ldots, x_K\}$ and each of them has $L_1 > 1$ available max-player actions $\{a_1, a_2, \ldots, a_{L_1}\} =: \mathcal{A}$, and $L_2 > 1$ min-player actions $\{b_1, b_2, \ldots, b_{L_2}\} =: \mathcal{B}$. Each state in $\mathcal{Y}_1 := \{y_{1,x,a,b} : \ \forall x \in \mathcal{X}, a \in \mathcal{A}, b \in \mathcal{B}\}$ and $\mathcal{Y}_2 := \{y_{2,x,a,b} : \ \forall x \in \mathcal{X}, a \in \mathcal{A}, b \in \mathcal{B}\}$ only has a single joint-action pair to choose. In total, there are $N := 3KL_1L_2$ state-joint-action pairs. For state $x \in \mathcal{X}$, by taking a joint-action (a, b) for $a \in \mathcal{A}, b \in \mathcal{B}$, it transitions to a state $y_{1,x,a,b} \in \mathcal{Y}_1$ with probability 1. For state $y_{1,x,a,b} \in \mathcal{Y}_1$, there is only a single joint-action for both players to choose from, which is the (a, b) pair that leads to this $y_{1,x,a,b}$. It then transitions to itself with probability $p_{\mathcal{G},x,a,b} \in (1/2, 1)$ and to a corresponding state $y_{2,x,a,b} \in \mathcal{Y}_2$ with probability $1 - p_{\mathcal{G},x,a,b}$. Note that $p_{\mathcal{G},x,a,b}$ can be different for different state-joint-action tuples. All states in $\mathcal{Y}_2$ are absorbing. The reward function is: for any $y_{1,x,a,b} \in \mathcal{Y}_1$, $R(y_{1,x,a,b}) = \iota_{\mathcal{G},x,a,b}$ for some $\iota_{\mathcal{G},x,a,b} \in [0, 1]$ (to be specified later); and for all other states, $R(s) = 0$. And the Q-function of the MGs can be computed as

$$Q_{\mathcal{G}}(x, a, b) = \frac{\gamma \iota_{\mathcal{G},x,a,b}}{1 - \gamma p_{\mathcal{G},x,a,b}}, \quad \forall\ (x, a, b) \in \mathcal{X} \times \mathcal{A} \times \mathcal{B}, \tag{71}$$

which is fully characterized by $p_{\mathcal{G},x,a,b}$ and $\iota_{\mathcal{G},x,a,b}$.

Transition Model Hypotheses of $\mathcal{G}$. We restrict $\gamma \in (1/2, 1)$. Let $p_0 = \gamma$ and $\alpha_1, \alpha_2 \in (0, 1)$. We consider $M+1$ possibilities of the transition models of $\mathcal{G}$, where $M := K[L_1(L_2 - 1)]$ — the null hypothesis is:

$$\mathcal{G}_1 : \begin{cases} p_{\mathcal{G}_1,x_k,a_1,b_1} = p_0 - \alpha_1, \quad \forall\ k \in [K], \\ p_{\mathcal{G}_1,x_k,a_l,b_1} = p_0 - 2\alpha_1, \quad \forall\ k \in [K], l \in [L_1]\backslash\{1\}, \\ p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}} = p_0, \quad \forall\ k \in [K], l_1 \in [L_1], l_2 \neq 1; \end{cases} \tag{72}$$

and for all $k \in [K]$, $l_1 \in [L_1]$, and $l_2 \in [L_2]\backslash\{1\}$ the M alternative hypotheses are:

$$\mathcal{G}_{k,l_1,l_2} : \begin{cases} p_{\mathcal{G}_{k,l_1,l_2},x_k,a_{l_1},b_{l_2}} = p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}} - \alpha_2 = p_0 - \alpha_2, \\ p_{\mathcal{G}_{k,l_1,l_2},x_k,a_{l_1'},b_{l_2}} = p_{\mathcal{G}_1,x_k,a_{l_1'},b_{l_2}} = p_0, \quad \forall l_1' \neq l_1, \\ p_{\mathcal{G}_{k,l_1,l_2},x_{k'},a_{l_1'},b_{l_2'}} = p_{\mathcal{G}_1,x_{k'},a_{l_1'},b_{l_2'}}, \quad \forall (k', l_1', l_2') \neq (k, l_1, l_2), \end{cases} \tag{73}$$

where $\alpha_1 = c'(1 - \gamma p_0)^2 \epsilon/\gamma$, $\alpha_2 = c(1 - \gamma p_0)^2 \epsilon/\gamma$ for some $\epsilon \in (0, 1)$ and absolute constants $c', c > 0$ to be determined later. Note that each alternative hypothesis only has one element in the transition model different from the null one.

Reward Functions. We define a class of $M+1$ reward functions $\mathfrak{R} = \{r_1\} \bigcup \{r_{k,l_1,l_2} : \forall k \in [K], l_1 \in [L_1], l_2 \in [L_2]\backslash\{1\}\}$, which is unknown to $\mathfrak{A}$ during sampling, and is defined as follows (recall that other than the value specified by $\iota_{\mathcal{G},x,a,b}$, rewards are all zero):

$$r_1 : \begin{cases} \iota_{\mathcal{G},x_k,a_1,b_1} = 1, \quad \forall\ k \in [K], \\ \iota_{\mathcal{G},x_k,a_l,b_1} = 1, \quad \forall\ k \in [K], l \in [L_1]\backslash\{1\}, \\ \iota_{\mathcal{G},x_k,a_{l_1},b_{l_2}} = 1, \quad \forall\ k \in [K], l_1 \in [L_1], l_2 \neq 1; \end{cases}$$

$$r_{k,l_1,l_2} : \begin{cases} \iota_{\mathcal{G},x_k,a_{l_1},b_{l_2}} = 1, \\ \iota_{\mathcal{G},x_k,a_{l_1'},b_{l_2}} = \frac{1-\gamma p_{\mathcal{G}_1,x_k,a_{l_1'},b_{l_2}}}{1-\gamma\left(p_{\mathcal{G}_1,x_k,a_{l_1'},b_{l_2}} - 2\alpha_2\right)}, \quad \forall l_1' \neq l_1, \\ \iota_{\mathcal{G},x_{k'},a_{l_1'},b_{l_2'}} = 1, \quad \forall (k', l_1', l_2') \neq (k, l_1, l_2), \end{cases}$$

for all $k \in [K]$, $l_1 \in [L_1]$, and $l_2 \in [L_2]\backslash\{1\}$.

By the construction above, if the reward function $r_m \in \mathfrak{R}$ is assigned to the corresponding transition model in $\mathcal{G}_m$, for either $m = 1$ or any $m = (k, l_1, l_2)$, then the corresponding Q-values become

$$\text{for}\ \ \mathcal{G}_1 : \begin{cases} Q_{\mathcal{G}_1}(x_k, a_1, b_1) = \frac{\gamma}{1-\gamma(p_0-\alpha_1)}, \quad \forall\ k \in [K], \\ Q_{\mathcal{G}_1}(x_k, a_l, b_1) = \frac{\gamma}{1-\gamma(p_0-2\alpha_1)}, \quad \forall\ k \in [K], l \in [L_1]\backslash\{1\}, \\ Q_{\mathcal{G}_1}(x_k, a_{l_1}, b_{l_2}) = \frac{\gamma}{1-\gamma p_0}, \quad \forall\ k \in [K], l_1 \in [L_1], l_2 \neq 1; \end{cases} \tag{74}$$

and $\forall k\in[K],\ l_1\in[L_1],\ l_2\in[L_2]\backslash\{1\}$,

$$\text{for}\ \ \mathcal{G}_{k,l_1,l_2}:\begin{cases}Q_{\mathcal{G}_{k,l_1,l_2}}(x_k,a_{l_1},b_{l_2})=\frac{\gamma}{1-\gamma\left(p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}-\alpha_2\right)},\\ Q_{\mathcal{G}_{k,l_1,l_2}}(x_k,a_{l_1'},b_{l_2})=\frac{\gamma}{1-\gamma\left(p_{\mathcal{G}_1,x_k,a_{l_1'},b_{l_2}}-2\alpha_2\right)},\quad \forall l_1'\neq l_1,\\ Q_{\mathcal{G}_{k,l_1,l_2}}(x_{k'},a_{l_1'},b_{l_2'})=\frac{\gamma}{1-\gamma p_{\mathcal{G}_1,x_{k'},a_{l_1'},b_{l_2'}}},\quad \forall(k',l_1',l_2')\neq(k,l_1,l_2).\end{cases}\tag{75}$$

We then select α_1 such that for $(l_1,l_2)\neq(1,1)$,

$$\begin{aligned}&\left|Q_{\mathcal{G}_1}(x_k,a_1,b_1)-Q_{\mathcal{G}_1}(x_k,a_{l_1},b_{l_2})\right|\\ &\qquad\geq\min\Big(\frac{\gamma}{1-\gamma p_0}-\frac{\gamma}{1-\gamma(p_0-\alpha_1)},\frac{\gamma}{1-\gamma(p_0-\alpha_1)}-\frac{\gamma}{1-\gamma(p_0-2\alpha_1)}\Big)\geq 20\epsilon\end{aligned}\tag{76}$$

and α_2 is selected such that $\alpha_2\geq 2\alpha_1$ and

$$48\epsilon\geq|Q_{\mathcal{G}_{k,l_1,l_2}}(x_k,a_{l_1},b_{l_2})-Q_{\mathcal{G}_{k,l_1,l_2}}(x_k,a_{l_1'},b_{l_2'})|\geq 20\epsilon\tag{77}$$

for all $(l_1',l_2')\neq(l_1,l_2)$. Moreover, we require that $p_0\in(1/2+2\alpha_1+2\alpha_2,1)$, $\alpha_2/(1-p_0)\in(0,1/2)$ and $\alpha_2/(p_0-2\alpha_1-2\alpha_2)\in(0,1/2)$. Hence, $\epsilon\leq\mathcal{O}(1/(1-\gamma))$.

In the sequel, we denote $\mathbb{E}_1$ and $\mathbb{P}_1$ to measure the expectation and probability of an event under the transition model hypothesis $\mathcal{G}_1$. Similarly, we denote $\mathbb{E}_{k,l_1,l_2}$ and $\mathbb{P}_{k,l_1,l_2}$ to measure the expectation and probability of an event under hypothesis $\mathcal{G}_{k,l_1,l_2}$. It is not hard to verify that in the above case (with r_m being assigned to $\mathcal{G}_m$ correspondingly, for $m=1$ or $m=(k,l_1,l_2)$), there is a unique NE policy pair under hypothesis $\mathcal{G}_1$: for $x\in\mathcal{X}$, $\mu_1^*(x)=a_1$ and $\nu_1^*(x)=b_1$; and that there is a unique NE policy under hypothesis $\mathcal{G}_{k,l_1,l_2}$ for all $l_1\in[L_1]$ and $l_2\in[L_2]\backslash\{1\}$: for $k'\neq k$, $\mu^*_{k,l_1,l_2}(x_{k'})=a_1$, $\nu^*_{k,l_1,l_2}(x_{k'})=b_1$, and $\mu^*_{k,l_1,l_2}(x_k)=a_{l_1}$, $\nu^*_{k,l_1,l_2}(x_k)=b_{l_2}$.

Moreover, one can verify that if any reward r_{k,l_1,l_2} with $k\in[K]$, $l_1\in[L_1]$, and $l_2\in[L_2]\backslash\{1\}$ (instead of r_1 as in (74)) is assigned to the transition model of $\mathcal{G}_1$, then the NE policy at x_k can never be the pure strategy $\mu_1^*(x_k)=a_{l_1}$, $\nu_1^*(x_k)=b_{l_2}$ (it can be some mixed NE policy). As a consequence, for algorithm $\mathfrak{A}$, after estimating the transition model of $\mathcal{G}_1$, if r_{k,l_1,l_2} is revealed, then it will output some ϵ-NE policy with probability greater than $1-\delta$; this ϵ-NE policy pair, which can be mixed strategies at x_k, should output the joint-action (a_{l_1},b_{l_2}) with a small probability, which is smaller than

$$\beta:=\frac{\frac{\gamma}{1-\gamma(p_0-\alpha_1)}-\frac{\gamma}{1-\gamma(p_0-2\alpha_2)}+\epsilon}{\frac{\gamma}{1-\gamma p_0}-\frac{\gamma}{1-\gamma(p_0-2\alpha_2)}}=1-\frac{\frac{\gamma}{1-\gamma p_0}-\frac{\gamma}{1-\gamma(p_0-\alpha_1)}-\epsilon}{\frac{\gamma}{1-\gamma p_0}-\frac{\gamma}{1-\gamma(p_0-2\alpha_2)}}\leq 1-\frac{19}{96}\leq 1-\frac{19\epsilon(1-\gamma p_0)}{\gamma},\tag{78}$$

(implying that $\epsilon\leq\gamma/[96(1-\gamma p_0)]$), where the first inequality is due to (76)-(77), and the last one follows by upper-bounding $\gamma/(1-\gamma p_0)-\gamma/[1-\gamma(p_0-2\alpha_2)]$ simply by $\gamma/(1-\gamma p_0)$. This is because otherwise, the value of the ϵ-NE policy at x_k, denoted by $V^*_{\mathcal{G}_1}(x_k)$ satisfies

$$V^*_{\mathcal{G}_1}(x_k)\geq\beta\cdot\frac{\gamma}{1-\gamma p_0}+(1-\beta)\cdot\frac{\gamma}{1-\gamma(p_0-2\alpha_2)}=\frac{\gamma}{1-\gamma(p_0-\alpha_1)}+\epsilon,\tag{79}$$

where the first inequality is because with reward r_{k,l_1,l_2} being assigned to model $\mathcal{G}_1$, at state x_k and with the joint-action (a_{l_1},b_{l_2}), the Q-value is $\gamma/(1-\gamma p_0)$, while the smallest Q-value at state x_k is $\gamma/[1-\gamma(p_0-2\alpha_2)]$; the last equation is due to the definition of β in (78). However, one can verify that the NE-value in this case lies in the range $[\gamma/(1-\gamma(p_0-2\alpha_1)), \gamma/(1-\gamma(p_0-\alpha_1))]$, by finding the minimax and maximin elements in the payoff matrix, i.e., the Q-value table at x_k (using Lemma 25). Thus, (79) contradicts the fact that this policy is an ϵ-NE policy (thus making $V^*_{\mathcal{G}_1}(x_k)$ ϵ-close to the NE-value). If we define the following events for every $k\in[K]$, $l_1\in[L_1]$, and $l_2\in[L_2]\backslash\{1\}$:

$$B_{k,l_1,l_2}=\Big\{\text{when fed with } r_{k,l_1,l_2}\in\mathfrak{R},\ \mathfrak{A} \text{ outputs } (\mu,\nu) \text{ s.t. at } x_k,\ \big(a_{l_1},b_{l_2}\big) \text{ is generated w.p.} \le\beta\Big\}, \tag{80}$$

then the above argument can be written as $\mathbb{P}_1\big(B_{k,l_1,l_2}\big)\ge 1-\delta$.

Now, we fix $\epsilon\in(0,\epsilon_0)$ and $\delta\in(0,\delta_0)$, where ϵ_0 and δ_0 will be determined later. Let

$$t^*=\frac{c_1}{(1-\gamma)^3\epsilon^2}\log\Big(\frac{1}{4\delta}\Big),$$

where $c_1>0$ is an absolute constant to be determined later. We also define T_{k,l_1,l_2} to be the number of samples that algorithm $\mathfrak{A}$ calls from the generative model with input state $y_{1,x_k,a_{l_1},b_{l_2}}$ till $\mathfrak{A}$ stops (these sample calls are not necessarily consecutive). Note that no reward information is used/revealed to the agent in this sampling process of $\mathfrak{A}$. For every $k\in[K]$, $l_1\in[L_1]$, and $l_2\in[L_2]\backslash\{1\}$, we define the following two events:

$$A_{k,l_1,l_2}=\{T_{k,l_1,l_2}\le 4t^*\}, \tag{81}$$

$$C_{k,l_1,l_2}=\Big\{S_{k,l_1,l_2}-p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}T_{k,l_1,l_2}\le\sqrt{2p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}(1-p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}})T_{k,l_1,l_2}\log(1/4\delta)}\Big\}, \tag{82}$$

where S_{k,l_1,l_2} is the number of transitions to itself in the T_{k,l_1,l_2} calls to the generative model with input state $y_{1,x_k,a_{l_1},b_{l_2}}$. For these events, we have the following lemmas.

Lemma 21 *For any $k\in[K]$, $l_1\in[L_1]$, and $l_2\in[L_2]\backslash\{1\}$, if $\mathbb{E}_1[T_{k,l_1,l_2}]\le t^*$, $\mathbb{P}_1(A_{k,l_1,l_2})>3/4$.*

Proof Notice that

$$t^*\ge\mathbb{E}_1[T_{k,l_1,l_2}]>4t^*\mathbb{P}_1(T_{k,l_1,l_2}>4t^*)=4t^*(1-\mathbb{P}_1(T_{k,l_1,l_2}\le 4t^*)).$$

Thus, $\mathbb{P}_1(A_{k,l_1,l_2})>3/4$. ■

Lemma 22 *For any $k\in[K]$, $l_1\in[L_1]$, and $l_2\in[L_2]\backslash\{1\}$, if $\delta<1/16$, $\mathbb{P}_1(C_{k,l_1,l_2})\ge 3/4$.*

Proof We denote *outcome* to be 1 if the transition from $y_{1,x_k,a_{l_1},b_{l_2}}$ ends up on itself; otherwise 0. By definition, the outcomes from state $y_{1,x_k,a_{l_1},b_{l_2}}$ are i.i.d. Bernoulli-$p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}$ random variables. Let $\epsilon:=\sqrt{2p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}(1-p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}})T_{k,l_1,l_2}\log(1/4\delta)}$. By Chernoff-

Hoeffding bound and $p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}} \geq p_0 - 2\alpha_1 > 1/2$, we have that

$$\begin{aligned}
&\mathbb{P}_1\bigg(S_{k,l} - p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}} T_{k,l_1,l_2} \leq \epsilon\bigg) \\
&\geq\ 1 - \exp\bigg(- \mathrm{KL}\bigg(p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}} + \frac{\epsilon}{T_{k,l_1,l_2}} \ \bigg|\bigg|\ p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}\bigg) \cdot T_{k,l_1,l_2}\bigg) \\
&\geq 1 - \exp\bigg(- \frac{\epsilon^2}{2p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}(1 - p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}})T_{k,l_1,l_2}}\bigg) \geq 1 - 4\delta. \qquad (83)
\end{aligned}$$

Additional application of $\delta < 1/16$ proves the lemma. ∎

Let $\delta_0 = 1/16$ and $\epsilon_0 = \gamma/[96(1-\gamma p_0)]$. Then, for $\delta \in (0, \delta_0)$ and $\epsilon \in (0, \epsilon_0)$, and with the transition model of $\mathcal{G}_1$ being input, by the argument after (80), we have $\mathbb{P}_1(B_{k,l_1,l_2}) \geq 1 - \delta \geq 1 - 1/16 \geq 3/4$, for all $k \in [K]$, $l_1 \in [L_1]$, and $l_2 \in [L_2]\backslash\{1\}$. Define the event $\mathcal{E}_{k,l_1,l_2} := A_{k,l_1,l_2} \cap B_{k,l_1,l_2} \cap C_{k,l_1,l_2}$. Combining Lemmas 21 and 22 and $\mathbb{P}_1(B_{k,l_1,l_2}) \geq 3/4$, we have that

$$\mathbb{P}_1(\mathcal{E}_{k,l_1,l_2}) > (3/4)^3 > 1/4, \quad \forall\ k \in [K],\ l_1 \in [L_1],\ l_2 \in [L_2]\backslash\{1\}, \qquad (84)$$

if $\mathbb{E}_1[T_{k,l_1,l_2}] \leq t^*$, $\delta \in (0, \delta_0)$ and $\epsilon \in (0, \epsilon_0)$. Next, we show that if the expectation of the number of samples in $\mathfrak{A}$ on any $y_{1,x_k,a_{l_1},b_{l_2}}$ is no greater than t^* under the hypothesis $\mathcal{G}_1$, then B_{k,l_1,l_2} occurs with probability greater than δ under the hypothesis $\mathcal{G}_{k,l_1,l_2}$.

Lemma 23 *Let* $\epsilon_0 = \min\Big\{\frac{\gamma}{96(1-\gamma p_0)},\ c'' \min\Big\{\frac{\gamma}{(1-\gamma p_0)^2}, \frac{1}{1-\gamma}\Big\}\Big\}$ *for some constant* $c'' > 0$. *For any* $k \in [K]$, $l_1 \in [L_1]$, *and* $l_2 \in [L_2]\backslash\{1\}$, *when* $\epsilon \in (0, \epsilon_0)$, *if* $\mathbb{E}_1[T_{k,l_1,l_2}] \leq t^*$, *then* $\mathbb{P}_{k,l_1,l_2}(B_{k,l_1,l_2}) \geq \delta$.

Proof Let W be the length-T_{k,l_1,l_2} random sequence of the next states by calling the generative model T_{k,l_1,l_2} times with the input state $y_{1,x_k,a_{l_1},b_{l_2}}$. To simplify notation, we represent W as a binary sequence where 1 represents the next state from $y_{1,x_k,a_{l_1},b_{l_2}}$ to itself and 0 otherwise. If $(l_1, l_2) \neq (1, 1)$ and $\mathcal{G} = \mathcal{G}_1$, W forms an i.i.d. Bernoulli-$p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}$ sequence; if $\mathcal{G} = \mathcal{G}_{k,l_1,l_2}$, this is an i.i.d Bernoulli-$p_{\mathcal{G}_{k,l_1,l_2},x_k,a_{l_1},b_{l_2}}$ sequence. We define the likelihood function $\mathcal{L}_{k,l_1,l_2}$ as

$$\forall w \in \{0,1\}^{T_{k,l_1,l_2}}: \quad \mathcal{L}_{k,l_1,l_2}(w) = \mathbb{P}_{k,l_1,l_2}[W = w] \quad \text{and} \quad \mathcal{L}_1(w) = \mathbb{P}_1[W = w].$$

Recall that the notation S_{k,l_1,l_2} denotes the total number of 1's in W. For convenience, let us denote

$$p_1 = p_{\mathcal{G}_1,x_k,a_{l_1},b_{l_2}}, \quad \text{and} \quad p_2 = p_{\mathcal{G}_{k,l_1,l_2},x_k,a_{l_1},b_{l_2}}.$$

Note that

$$p_1 - p_2 = \alpha_2.$$

To additionally simplify the notation, we define $T = T_{k,l_1,l_2}$ and $S = S_{k,l_1,l_2}$. With these new notations, we compute $\mathcal{L}_{k,l_1,l_2}(W)/\mathcal{L}_1(W)$ as follows

$$\begin{aligned}\frac{\mathcal{L}_{k,l_1,l_2}(W)}{\mathcal{L}_1(W)} &= \frac{(p_2)^S(1-p_2)^{T-S}}{(p_1)^S(1-p_1)^{T-S}} = \left(1+\frac{\alpha_2}{p_1}\right)^S\left(1-\frac{\alpha_2}{1-p_1}\right)^{T-S} \\ &= \left(1+\frac{\alpha_2}{p_1}\right)^S\left(1-\frac{\alpha_2}{1-p_1}\right)^{S\frac{1-p_1}{p_1}}\left(1-\frac{\alpha_2}{1-p_1}\right)^{T-S/p_1}. \end{aligned} \tag{85}$$

Note that $p_0-2\alpha_1 \le p_1 \le p_0$. By our choice of p_0, α_1, α_2, and ϵ, it holds that $\alpha_2/(1-p_1)\in(0,1/2)$ and $\alpha_2/p_1\in(0,1/2)$. With the fact that $\log(1-u)\ge -u-u^2$ for $u\in[0,1/2]$ and $\exp(-u)\ge 1-u$ for $u\in[0,1]$, we have that

$$\left(1-\frac{\alpha_2}{1-p_1}\right)^{\frac{1-p_1}{p_1}} \ge \exp\left(\frac{1-p_1}{p_1}\left(-\frac{\alpha_2}{1-p_1}-(\frac{\alpha_2}{1-p_1})^2\right)\right) \ge \left(1-\frac{\alpha_2}{p_1}\right)\left(1-\frac{\alpha_2^2}{p_1(1-p_1)}\right). \tag{86}$$

Thus

$$\begin{aligned}\frac{\mathcal{L}_{k,l_1,l_2}(W)}{\mathcal{L}_1(W)} &\ge \left(1-\frac{\alpha_2^2}{p_1^2}\right)^S\left(1-\frac{\alpha_2^2}{p_1(1-p_1)}\right)^S\left(1-\frac{\alpha_2}{1-p_1}\right)^{T-S/p_1} && (87)\\ &\ge \left(1-\frac{\alpha_2^2}{p_1^2}\right)^T\left(1-\frac{\alpha_2^2}{p_1(1-p_1)}\right)^T\left(1-\frac{\alpha_2}{1-p_1}\right)^{T-S/p_1} && (88)\end{aligned}$$

due to $S\le T$. Next, we proceed on the event $\mathcal{E}_{k,l_1,l_2}$. By definition, if $\mathcal{E}_{k,l_1,l_2}$ occurs, event A_{k,l_1,l_2} has occurred. Using $\log(1-u)\ge -2u$ for $u\in[0,1/2]$, it follows that

$$\left(1-\frac{\alpha_2^2}{p_1^2}\right)^T \ge \left(1-\frac{\alpha_2^2}{p_1^2}\right)^{4t^*} \ge \exp\left(-8t^*\frac{\alpha_2^2}{p_1^2}\right) \ge (4\delta)^{128c^2c_1},$$

where we use the fact that

$$\begin{aligned} t^*\cdot\frac{\alpha_2^2}{p_1^2} &= \frac{c_1}{(1-\gamma)^3\epsilon^2}\log\left(\frac{1}{4\delta}\right)\cdot\frac{c^2(1-\gamma p_0)^4\epsilon^2}{\gamma^2p_1^2} \le \frac{c_1}{(1-\gamma)^3}\log\left(\frac{1}{4\delta}\right)\cdot\frac{c^2(1-\gamma p_0)^4}{\gamma^2p_1^2}\\ &\le 16c_1c^2(1-\gamma)\cdot\log(1/4\delta).\end{aligned}$$

Using $\log(1-u)\ge -2u$ for $u\in[0,1/2]$, we also have that

$$\left(1-\frac{\alpha_2^2}{p_1(1-p_1)}\right)^T \ge \left(1-\frac{\alpha_2^2}{p_1(1-p_1)}\right)^{4t^*} \ge \exp\left(-8t^*\frac{\alpha_2^2}{p_1(1-p_1)}\right) \ge (4\delta)^{64c^2c_1},$$

where we use

$$\begin{aligned} t^*\cdot\frac{\alpha_2^2}{p_1(1-p_1)} &= \frac{c_1}{(1-\gamma)^3\epsilon^2}\log\left(\frac{1}{4\delta}\right)\cdot\frac{c^2(1-\gamma p_0)^4\epsilon^2}{\gamma^2p_1(1-p_1)}\\ &\le \frac{c_1}{(1-\gamma)^3}\log\left(\frac{1}{4\delta}\right)\cdot\frac{c^2(1-\gamma p_0)^4}{\gamma^2(p_1)(1-p_0)} \le 8c_1c^2\cdot\log(1/4\delta).\end{aligned}$$

Further, we have that when $\mathcal{E}_{k,l_1,l_2}$ occurs, C_{k,l_1,l_2} also occurs. Therefore,

$$\begin{aligned}\left(1-\frac{\alpha_2}{1-p_1}\right)^{T-S/p_1} &\geq \left(1-\frac{\alpha_2}{1-p_1}\right)^{\sqrt{\frac{1-p_1}{p_1}T\log(1/4\delta)}} \geq \left(1-\frac{\alpha_2}{1-p_1}\right)^{\sqrt{\frac{1-p_1}{p_1}4t^*\log(1/4\delta)}} \\ &\geq \exp\left(-\sqrt{16\frac{\alpha_2^2}{p_1(1-p_1)}t^*\log(1/4\delta)}\right) \geq (4\delta)^{\sqrt{16c_1c^2}}.\end{aligned}$$

By taking c_1 small enough, e.g., $c_1 = 10^{-5}c^{-2}$, we have $\mathcal{L}_{k,l_1,l_2}(W)/\mathcal{L}_1(W) \geq 4\delta$. Note that by (73), the probability measure of the whole sample sequence under the two hypotheses $\mathcal{G}_1$ and $\mathcal{G}_{k,l_1,l_2}$ only differ at (k, l_1, l_2). By a change of measure, we deduce that

$$\mathbb{P}_{k,l_1,l_2}(B_{k,l_1,l_2}) \geq \mathbb{P}_{k,l_1,l_2}(\mathcal{E}_{k,l_1,l_2}) = \mathbb{E}_{k,l_1,l_2}[\mathbf{1}_{\mathcal{E}_{k,l_1,l_2}}] = \mathbb{E}_1\left[\frac{\mathcal{L}_{k,l_1,l_2}(W)}{\mathcal{L}_1(W)}\mathbf{1}_{\mathcal{E}_{k,l_1,l_2}}\right] \geq 4\delta\cdot 1/4 = \delta, \tag{89}$$

which completes the proof. ■

If $\mathfrak{A}$ is an (ϵ, δ)-correct reward-agnostic algorithm, then under transition model hypothesis $\mathcal{G}_{k,l_1,l_2}$, when fed with r_{k,l_1,l_2}, it produces an ϵ-NE policy pair (μ, ν) with probability at least $1-\delta$. At state x_k, this ϵ-NE policy should generate the joint-action (a_{l_1}, b_{l_2}) with a high probability. To see this, note that now (a_{l_1}, b_{l_2}) is the unique NE strategy at state x_k, which is a pure strategy. By Lemma 26, $(\mu(\cdot\,|\,x_k), \mathbb{1}_{b=b_{l_2}})$ is an 2ϵ-NE strategy at x_k. Let $\zeta \in [0, 1]$ denote the probability of choosing a_{l_1}, i.e., $\zeta = \mu(a_{l_1}\,|\,x_k)$. Then, the value at x_k under $(\mu(\cdot\,|\,x_k), \mathbb{1}_{b=b_{l_2}})$, denoted by $V_{\mu,b_{l_2}}(x_k)$, is

$$V_{\mu,b_{l_2}}(x_k) = \zeta\cdot\frac{\gamma}{1-\gamma(p_0-\alpha_2)} + (1-\zeta)\cdot\frac{\gamma}{1-\gamma(p_0-2\alpha_2)} \leq \frac{\gamma}{1-\gamma(p_0-\alpha_2)},$$

which, by the 2ϵ-NE property, should satisfy

$$V_{\mu,b_{l_2}}(x_k) \geq \frac{\gamma}{1-\gamma(p_0-\alpha_2)} - 2\epsilon \Longrightarrow \zeta \geq 1 - \frac{2\epsilon}{\frac{\gamma}{1-\gamma(p_0-\alpha_2)} - \frac{\gamma}{1-\gamma(p_0-2\alpha_2)}} \geq 1 - \frac{2\epsilon}{20\epsilon} = \frac{9}{10}.$$

Similarly, let $\xi = \nu(b_{l_2}\,|\,x_k)$, we have

$$\begin{aligned}V_{a_{l_1},\nu}(x_k) &\geq \xi\cdot\frac{\gamma}{1-\gamma(p_0-\alpha_2)} + (1-\xi)\cdot\left[\frac{\gamma}{1-\gamma(p_0-\alpha_2)} + 20\epsilon\right] \\ &= \frac{\gamma}{1-\gamma(p_0-\alpha_2)} + 20(1-\xi)\epsilon,\end{aligned}$$

where the inequality is due to (77). As $(\mathbb{1}_{a=a_{l_1}}, \nu(\cdot\,|\,x_k))$ is an 2ϵ-NE at x_k, we have $V_{a_{l_1},\nu}(x_k) \leq \gamma/[1-\gamma(p_0-\alpha_2)] + 2\epsilon$, leading to $\xi \geq 9/10$. Thus, for the ϵ-NE (μ, ν), the probability of generating (a_{l_1}, b_{l_2}) is at least $\zeta\cdot\xi \geq 81/100$.

Hence, recalling the definition in (80) and the fact that $\beta \leq 1 - 19/96 < 81/100$, we have $\mathbb{P}_{k,l_1,l_2}\big(B_{k,l_1,l_2}\big) < \delta$ for all $k \in [K]$, $l_1 \in [L_1]$, and $l_2 \in [L_2]\backslash\{1\}$. From Lemma 23, this does not happen unless $\mathbb{E}_1[T_{k,l_1,l_2}] > t^*$ for all $k \in [K]$, $l_1 \in [L_1]$, and $l_2 \in [L_2]\backslash\{1\}$. By linearity of expectation, the expected number of samples required by $\mathfrak{A}$ under hypothesis

$\mathcal{G}_1$ is at least $K[L_1(L_2-1)]t^* = \Omega\left(\frac{N}{(1-\gamma)^3\epsilon^2}\log(1/\delta)\right)$, which proves the lower bound for finding the ϵ-NE policy.

On the lower bound for finding ϵ-approximate NE value, the hard cases above can also be used. In fact, suppose some algorithm $\mathfrak{A}$ returns some $\widehat{Q}$ such that $\|\widehat{Q}-Q^*\|_\infty \leq \epsilon/4$ with probability at least $1-\delta$, then it can identify the pure NE strategy as described in the paragraph before (80) for the Q-values given in (74)-(75) (when reward r_m is assigned to transition model of $\mathcal{G}_m$ correspondingly), under our choices of the parameters. This can be done by solving for the NE of the corresponding $\widehat{Q}$. Moreover, when r_{k,l_1,l_2} is assigned to $\mathcal{G}_1$ (instead of r_1 as in (74)), this procedure of solving the NE policy for $\widehat{Q}$ will also output some policy that makes B_{k,l_1,l_2} in (80) hold with $\mathbb{P}_1(B_{k,l_1,l_2}) \geq 1-\delta$, following similar arguments around (79). Indeed, otherwise, if this procedure outputs $\mu(x_k)=a_{l_1}, \nu(x_k)=b_{l_2}$ with a high probability, then the NE value under payoff matrix $\widehat{Q}(x_k,\cdot,\cdot)$ will be at least $\epsilon/2$ away from the NE value under payoff matrix $Q^*(x_k,\cdot,\cdot)$, due to our choice of α_1. However, as one-step of the max min operation onto $\widehat{Q}(x_k,\cdot,\cdot)$ (and $Q^*(x_k,\cdot,\cdot)$) is non-expansive, the NE values under these two payoff matrices should differ no greater than $\epsilon/4$, as $\|\widehat{Q}-Q^*\|_\infty \leq \epsilon/4$. This shows $\mathbb{P}_1(B_{k,l_1,l_2}) \geq 1-\delta$. Then, using almost identical arguments as above, we obtain a lower bound of the same order, and thus prove Theorem 4.

Appendix B. Auxiliary Results

B.1 A Smooth Planning Oracle

We now show that solving the regularized matrix game induced by $\widehat{Q}^*$, see (6), leads to a smooth **Planning Oracle** with certain smoothness coefficient C (see Definition 7).

Lemma 24 *Suppose that the nonnegative regularizers Ω_i for $i=1,2$ in (6) are twice continuously differentiable, strongly convex, and bounded over the simplex. Suppose that for each $s\in\mathcal{S}$, the solution policy pair $(\widehat{\mu}(\cdot\,|\,s),\widehat{\nu}(\cdot\,|\,s))$ of (6) with $\tau_1=\tau_2=(1-\gamma)^2\epsilon>0$ lies in the relative interior of the simplexes $\Delta(\mathcal{A})$ and $\Delta(\mathcal{B})$, respectively. Then, $(\widehat{\mu},\widehat{\nu})$ is smooth with respect to $\widehat{Q}^*$, namely, this **Planning Oracle** follows Definition 7, with some constant $C=\textbf{poly}(|\mathcal{A}|,|\mathcal{B}|,|\mathcal{S}|,1/\epsilon,1/(1-\gamma))$, and meanwhile $\|\widehat{V}^{\widehat{\mu},*}-\widehat{V}^*\|_\infty \leq \mathcal{O}((1-\gamma)\epsilon)$, $\|\widehat{V}^{*,\widehat{\nu}}-\widehat{V}^*\|_\infty \leq \mathcal{O}((1-\gamma)\epsilon)$, namely, ϵ_{opt} in Theorem 8 satisfies $\epsilon_{opt}\leq\mathcal{O}((1-\gamma)\epsilon)$.*

Proof Let $Q_s := \widehat{Q}^*(s,\cdot,\cdot)\in\mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ denote the payoff matrix of the game at state s. Note that $Q_s\in[0,(1-\gamma)^{-1}]^{|\mathcal{A}|\times|\mathcal{B}|}$, $u\in[0,1]^{|\mathcal{A}|}$ and $\vartheta\in[0,1]^{|\mathcal{B}|}$. Also, note that by the simplex constraints on u,ϑ, there are $|\mathcal{A}|-1$ and $|\mathcal{B}|-1$ free variables, and the last dimension can be represented as $1-\sum_{i=1}^{|\mathcal{A}|-1}u(a_i)$, where we use a_i to denote the i-th element in $\mathcal{A}=\{a_1,a_2,\cdots,a_{|\mathcal{A}|}\}$. Thus, we introduce new vectors $\widetilde{u}=[u(a_1),u(a_2),\cdots,u(a_{|\mathcal{A}|-1})]^\top$ and $\widetilde{\vartheta}=[\vartheta(a_1),\vartheta(a_2),\cdots,\vartheta(a_{|\mathcal{A}|-1})]^\top$ of dimensions $\mathbb{R}^{|\mathcal{A}|-1}$ and $\mathbb{R}^{|\mathcal{B}|-1}$, respectively. As the solution to (6) lies in the relative interior of the simplex, we know that $1-\sum_{i=1}^{|\mathcal{A}|-1}u(a_i)>0$ and $1-\sum_{i=1}^{|\mathcal{B}|-1}\vartheta(b_i)>0$. We can then redefine the objective in (6) as

$$f(\widetilde{u},\widetilde{\vartheta}) := \Lambda_1(\widetilde{u})^\top Q_s\Lambda_2(\widetilde{\vartheta})-\tau_1\Omega_1(\Lambda_1(\widetilde{u}))+\tau_2\Omega_2(\Lambda_2(\widetilde{\vartheta})) \tag{90}$$

where $u = \Lambda_1(\widetilde{u}) = \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix} \widetilde{u} + \boldsymbol{e}_{|\mathcal{A}|}$ and $\vartheta = \Lambda_2(\widetilde{\vartheta}) = \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix} \widetilde{\vartheta} + \boldsymbol{e}_{|\mathcal{B}|}$, $\mathbf{1}$ denotes the all-one vector of proper dimension, and $\boldsymbol{e}_i$ denotes the vector of proper dimension whose i-th element is one and all other elements are zero.

Since the solution lies in the relative interior of $\Delta(\mathcal{A})$ and $\Delta(\mathcal{B})$, by first-order optimality, we have that for each $s \in \mathcal{S}$

$$\nabla_{\widetilde{u}} f(\widetilde{u}, \widetilde{\vartheta}) = -\tau_1 \nabla_{\widetilde{u}} \Omega_1(\Lambda_1(\widetilde{u})) + \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} Q_s \Lambda_2(\widetilde{\vartheta}) = 0, \tag{91}$$

$$\nabla_{\widetilde{\vartheta}} f(\widetilde{u}, \widetilde{\vartheta}) = \tau_2 \nabla_{\widetilde{\vartheta}} \Omega_2(\Lambda_2(\widetilde{\vartheta})) + \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} Q_s^\top \Lambda_1(\widetilde{u}) = 0, \tag{92}$$

whose solution is unique since (90) is still a strongly-convex-strongly-concave minimax problem. In particular, note that by the chain rule, the Hessians of f are $\nabla^2_{\widetilde{u}} f(\widetilde{u}, \widetilde{\vartheta}) = \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} \nabla^2_u g(u, \vartheta) \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix}$ and $\nabla^2_{\widetilde{\vartheta}} f(\widetilde{u}, \widetilde{\vartheta}) = \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} \nabla^2_\vartheta g(u, \vartheta) \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix}$, where

$$g(u, \vartheta) := u^\top Q_s \vartheta - \tau_1 \Omega_1(u) + \tau_2 \Omega_2(\vartheta)$$

is the original objective used in (6). Let $\eta_i > 0$ be the strong-convexity coefficient for Ω_i, then, we have $\nabla^2_{\widetilde{u}} f(\widetilde{u}, \widetilde{\vartheta}) \preceq -\tau_1 \eta_1 \cdot I$ and $\nabla^2_{\widetilde{\vartheta}} f(\widetilde{u}, \widetilde{\vartheta}) \succeq \tau_2 \eta_2 \cdot I$, since for any vector $x \in \mathbb{R}^{|\mathcal{A}|-1}$ (or $x \in \mathbb{R}^{|\mathcal{B}|-1}$) that is not $\mathbf{0}$, $\begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix} x$ is not $\mathbf{0}$.

Define a function $F : \mathbb{R}^{|\mathcal{A}|-1} \times \mathbb{R}^{|\mathcal{B}|-1} \times \mathbb{R}^{|\mathcal{A}||\mathcal{B}|} \to \mathbb{R}^{|\mathcal{A}|+|\mathcal{B}|-2}$ as follows, such that (91)-(92) is equivalent to

$$F\big(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)\big) := \begin{bmatrix} \tau_1 \nabla_{\widetilde{u}} \Omega_1(\Lambda_1(\widetilde{u})) - \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} Q_s \Lambda_2(\widetilde{\vartheta}) \\ \tau_2 \nabla_{\widetilde{\vartheta}} \Omega_2(\Lambda_2(\widetilde{\vartheta})) + \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} Q_s^\top \Lambda_1(\widetilde{u}) \end{bmatrix} = 0.$$

As the solution to (91)-(92) lies in the relative interior of the simplexes, for any choice of $Q_s \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{B}|}$ (not just $[0, (1-\gamma)^{-1}]^{|\mathcal{A}| \times |\mathcal{B}|}$), the domain of F can be specified as $\Delta^o(\mathcal{A}) \times \Delta^o(\mathcal{B}) \times \mathcal{Q}^o$, where $\Delta^o(\mathcal{A})$ and $\Delta^o(\mathcal{B})$ denote the sets of $\widetilde{u}$ and $\widetilde{\vartheta}$ whose corresponding u and ϑ lie in the interiors of $\Delta(\mathcal{A})$ and $\Delta(\mathcal{B})$, respectively, and $\mathcal{Q}^o \subset \mathbb{R}^{|\mathcal{A}||\mathcal{B}|}$ denotes some open set that contains $[0, (1-\gamma)^{-1}]^{|\mathcal{A}||\mathcal{B}|}$.

Notice that the Jacobian of F with respect to $[\widetilde{u}^\top \;\; \widetilde{\vartheta}^\top]^\top$ is

$$M\big(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)\big) := \begin{bmatrix} \frac{\partial F}{\partial \widetilde{u}} & \frac{\partial F}{\partial \widetilde{\vartheta}} \end{bmatrix} = \begin{bmatrix} \tau_1 \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} \nabla^2_u \Omega_1(\Lambda_1(\widetilde{u})) \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix} & -\begin{bmatrix} I & -\mathbf{1} \end{bmatrix} Q_s \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix} \\ \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} Q_s^\top \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix} & \tau_2 \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} \nabla^2_\vartheta \Omega_2(\Lambda_2(\widetilde{\vartheta})) \begin{bmatrix} I \\ -\mathbf{1}^\top \end{bmatrix} \end{bmatrix}, \tag{93}$$

which is always invertible for any point in $\Delta^o(\mathcal{A}) \times \Delta^o(\mathcal{B}) \times \mathcal{Q}^o$. This is because Ω_i are strongly convex, and thus the real parts of the eigenvalues of the matrix, which are the eigenvalues of $(M + M^\top)/2$, are always positive and uniformly lower bounded. Specifically, we have

$$\begin{aligned} &\min_i \lambda_i\big(M(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)) + M^\top(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s))\big) \\ &\qquad \geq 2 \min\{\tau_1 \eta_1, \;\tau_2 \eta_2\} = 2(1-\gamma)^2 \min\{\eta_1, \;\eta_2\} \cdot \epsilon > 0, \end{aligned}$$

with $\lambda_i(\cdot)$ being the i-th largest eigenvalues of the corresponding matrix. This further implies that for any $(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)) \in \Delta^o(\mathcal{A}) \times \Delta^o(\mathcal{B}) \times \mathcal{Q}^o$,

$$\left\|M(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s))^{-1}\right\|_2 = \frac{1}{\min_i \ \sigma_i(M(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)))} \tag{94}$$

$$\leq \frac{2}{\min_i \ \lambda_i(M(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)) + M^\top(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)))} \leq \frac{1}{\min\{\eta_1, \ \eta_2\}(1-\gamma)^2 \cdot \epsilon}, \tag{95}$$

where $\sigma_i(\cdot)$ is the i-th largest singular value of the corresponding matrix.

By the implicit function theorem (Krantz and Parks, 2012), for any point that solves $F(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s)) = 0$, since $M(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s))$ is invertible, there exists a neighborhood $U \subseteq \Delta^o(\mathcal{A})$, $V \subseteq \Delta^o(\mathcal{B})$, and $W \subseteq \mathcal{Q}^o$ around it, such that $[\widetilde{u}^\top \quad \widetilde{\vartheta}^\top]^\top \in U \times V$ is a unique function of $\mathrm{vec}(Q_s)$ for all $\mathrm{vec}(Q_s) \in W$, and

$$\frac{\partial[\widetilde{u}^\top \ \widetilde{\vartheta}^\top]^\top}{\partial \mathrm{vec}(Q_s)} = -\begin{bmatrix} \frac{\partial F}{\partial \widetilde{u}} & \frac{\partial F}{\partial \widetilde{\vartheta}} \end{bmatrix}^{-1} \cdot \frac{\partial F}{\partial \mathrm{vec}(Q_s)} = -M(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s))^{-1} \cdot \begin{bmatrix} -\Lambda_2(\widetilde{\vartheta})^\top \otimes \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} \\ \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} \otimes \Lambda_1(\widetilde{u})^\top \end{bmatrix},$$

where $\otimes$ denotes the Kronecker product. Thus, we have

$$\begin{aligned} \left\|\frac{\partial[\widetilde{u}^\top \ \widetilde{\vartheta}^\top]^\top}{\partial \mathrm{vec}(Q_s)}\right\|_2 &\leq \left\|M(\widetilde{u}, \widetilde{\vartheta}, \mathrm{vec}(Q_s))^{-1}\right\|_2 \cdot \left\|\frac{\partial F}{\partial \mathrm{vec}(Q_s)}\right\|_2 \\ &\leq \frac{\sqrt{(|\mathcal{A}| + |\mathcal{B}| - 2)|\mathcal{A}||\mathcal{B}|}}{\min\{\eta_1, \ \eta_2\}(1-\gamma)^2 \cdot \epsilon} \cdot \left\|\frac{\partial F}{\partial \mathrm{vec}(Q_s)}\right\|_\infty \leq \frac{2(|\mathcal{A}|\sqrt{|\mathcal{B}|} + |\mathcal{B}|\sqrt{|\mathcal{A}|})}{\min\{\eta_1, \ \eta_2\}(1-\gamma)^2 \cdot \epsilon}, \end{aligned}$$

where we have used (94), the fact that for matrix $A \in \mathbb{R}^{m \times n}$, $\|A\|_2 \leq \sqrt{mn}\|A\|_\infty$, and the fact that

$$\left\|\frac{\partial F}{\partial \mathrm{vec}(Q_s)}\right\|_\infty = 2 \cdot \max\left\{\|\Lambda_2(\widetilde{\vartheta})\|_1, \ \|\Lambda_1(\widetilde{u})\|_1\right\} = 2.$$

Notice that this is a uniform bound on the gradient of the implicit function, at any point in $\Delta^o(\mathcal{A}) \times \Delta^o(\mathcal{B}) \times \mathcal{Q}^o$, which together with the mean-value theorem leads to

$$\left\|[\widetilde{u}_1^\top \ \widetilde{\vartheta}_1^\top] - [\widetilde{u}_2^\top \ \widetilde{\vartheta}_2^\top]\right\|_2 \leq \frac{2(|\mathcal{A}|\sqrt{|\mathcal{B}|} + |\mathcal{B}|\sqrt{|\mathcal{A}|})}{\min\{\eta_1, \ \eta_2\}(1-\gamma)^2 \cdot \epsilon} \cdot \left\|\mathrm{vec}(Q_{s,1}) - \mathrm{vec}(Q_{s,2})\right\|_2,$$

where the pair $(\widetilde{u}_i, \widetilde{\vartheta}_i)$ is the unique solution of $F = 0$ corresponding to $Q_{s,i}$. By the equivalence of norms and considering all $s \in \mathcal{S}$, we can find some constant C (which depends on $|\mathcal{A}|$, $|\mathcal{B}|$, $|\mathcal{S}|$, as well as $1/\epsilon$ and $1/(1-\gamma)$ polynomially) as the smooth coefficient, and this completes the first argument of the result.

Now, it suffices to prove that the obtained solution $(\widehat{\mu}, \widehat{\nu})$ with parameter $\tau_1 = \tau_2 = (1-\gamma)^2\epsilon$ also leads to small ϵ_{opt}. Let $D_i > 0$ denotes the upper bound of the regularizer Ω_i

over the simplex. Then, we have that for any $s \in \mathcal{S}$

$$
\begin{aligned}
0 \leq \widehat{V}^*(s) - \widehat{V}^{\widehat{\mu},*}(s) &= \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^{\widehat{\mu},*}(s,a,b)\big] \\
&= \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] \\
&\quad + \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^{\widehat{\mu},*}(s,a,b)\big] \\
&\leq \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] + \gamma\|\widehat{V}^* - \widehat{V}^{\widehat{\mu},*}\|_\infty \qquad (96) \\
&\leq \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \Bigg(\min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \tau_1\Omega_1(\widehat{\mu}(\cdot\,|\,s)) \\
&\quad + \tau_2\Omega_2(\vartheta)\Bigg) + \tau_2 D_2 + \gamma\|\widehat{V}^* - \widehat{V}^{\widehat{\mu},*}\|_\infty \qquad (97) \\
&\leq \min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \Bigg(\min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] - \tau_1\Omega_1(\widehat{\mu}^*(\cdot\,|\,s)) \\
&\quad + \tau_2\Omega_2(\vartheta)\Bigg) + \tau_2 D_2 + \gamma\|\widehat{V}^* - \widehat{V}^{\widehat{\mu},*}\|_\infty \qquad (98) \\
&\leq \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\widetilde{\vartheta}}\big[\widehat{Q}^*(s,a,b)\big] - \Bigg(\min_{\vartheta\in\Delta(\mathcal{B})} \mathbb{E}_{a\sim\widehat{\mu}^*(\cdot\,|\,s),b\sim\vartheta}\big[\widehat{Q}^*(s,a,b)\big] + \tau_2\Omega_2(\vartheta)\Bigg) \\
&\quad + \tau_1 D_1 + \tau_2 D_2 + \gamma\|\widehat{V}^* - \widehat{V}^{\widehat{\mu},*}\|_\infty \qquad (99) \\
&\leq \tau_1 D_1 + \tau_2 D_2 + \gamma\|\widehat{V}^* - \widehat{V}^{\widehat{\mu},*}\|_\infty \qquad (100)
\end{aligned}
$$

where $(\widehat{\mu}^*, \widehat{\nu}^*)$ denotes a Nash equilibrium policy in the empirical model $\widehat{G}$, with $\widehat{V}^* = \widehat{V}^{\widehat{\mu}^*,\widehat{\nu}^*}$, (96) uses Bellman equation to relate Q-function and V-function, (97) uses the boundedness of Ω_2, (98) uses (6), and in (99) $\widetilde{\vartheta}$ denotes the argmin of $\vartheta \in \Delta(\mathcal{B})$ in the second term in (98). By the choices of $\tau_1 = \tau_2 = (1-\gamma)^2\epsilon$ and (100), we have that $\|\widehat{V}^* - \widehat{V}^{\widehat{\mu},*}\|_\infty \leq \mathcal{O}(\max\{\tau_1,\tau_2\}/(1-\gamma)) = \mathcal{O}((1-\gamma)\epsilon)$. The proof for $\|\widehat{V}^{*,\widehat{\nu}} - \widehat{V}^*\|_\infty \leq \mathcal{O}((1-\gamma)\epsilon)$ is symmetric and analogous. This completes the proof. ■

To ensure that the solution $(\widehat{\mu}(\cdot\,|\,s), \widehat{\nu}(\cdot\,|\,s))$ of (6) lies in the relative interior of the simplexes, the common choice of *steep* regularizers will suffice (Mertikopoulos and Sandholm, 2016). The steep regularizer means that for any u (resp. ϑ) on the boundary of the simplex $\Delta(\mathcal{A})$ (resp. $\Delta(\mathcal{B})$), and for every interior sequence $u_n \to u$ (resp. $\vartheta_n \to \vartheta$) that approaches it, it holds that $\big\|\frac{d\Omega_1(u)}{du}\big|_{u=u_n}\big\|_2 \to \infty$ (resp. $\big\|\frac{d\Omega_2(\vartheta)}{d\vartheta}\big|_{\vartheta=\vartheta_n}\big\|_2 \to \infty$). This way, the optimizer is not on the boundary of the simplexes. Examples of steep regularizers in Lemma 24 include the commonly used negative entropy, Tsallis entropy and Rényi entropy with certain parameters; see Mertikopoulos and Sandholm (2016) for more discussions. Also note that they are bounded over simplex for standard choices of the parameters, and thus satisfy the conditions in our Lemma 24.

B.2 Properties of (ϵ-)NE in Zero-Sum Matrix Games

Now we establish several properties of the (ϵ-)NE strategies in zero-sum matrix games, which have been used in the proof in §A.2.

Lemma 25 (NE Value Range) *Consider a two-player zero-sum matrix game $\mathcal{M}$ with the action spaces $\mathcal{A}$ and $\mathcal{B}$, and the payoff matrix $M \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ with $= \{m_{ij}\}_{i\in[|\mathcal{A}|],j\in[|\mathcal{B}|]}$ for the maximizer (agent-1). Then, the NE value of the game V^* is bounded between the* maximin *and* minimax *elements in M, i.e.,*

$$\max_i \min_j \ m_{ij} \leq V^* \leq \min_j \max_i \ m_{ij}.$$

Proof Note that

$$\max_{u\in\Delta(\mathcal{A})} \min_j \ u^\top M \boldsymbol{e}_j = \max_{u\in\Delta(\mathcal{A})} \min_{\vartheta\in\Delta(\mathcal{B})} \ u^\top M\vartheta = V^* = \min_{\vartheta\in\Delta(\mathcal{B})} \max_{u\in\Delta(\mathcal{A})} \ u^\top M\vartheta = \min_{\vartheta\in\Delta(\mathcal{B})} \max_i \ \boldsymbol{e}_i^\top M\vartheta,$$

where $\boldsymbol{e}_i$ denote the all-zero vector except a single 1 at element i, with proper dimensions. Also, notice that

$$\min_{\vartheta\in\Delta(\mathcal{B})} \max_i \ \boldsymbol{e}_i^\top M\vartheta \leq \min_j \max_i \ \boldsymbol{e}_i^\top M\boldsymbol{e}_j = \min_j \max_i \ m_{ij},$$

where the inequality is due to that $\boldsymbol{e}_i \in \Delta(\mathcal{B})$ and the min on the right is taken over a smaller set, thus has a larger value. This proves the right-hand side of the inequality. Proof for the other side is analogous. ■

Lemma 26 (ϵ-NE Strategy Interchangeability) *Consider the game as above in Lemma 25. Let $u_1, u_2 \in \Delta(\mathcal{A})$ and $\vartheta_1, \vartheta_2 \in \Delta(\mathcal{B})$ be strategies such that (u_1, ϑ_1) is a Nash equilibrium strategy, and (u_2, ϑ_2) is an ϵ-NE strategy. Then, both (u_1, ϑ_2) and (u_2, ϑ_1) are 2ϵ-NE strategy pairs.*

Proof Let $V(u, \vartheta) := u^\top M\vartheta$ denote the value under any strategy pair (u, ϑ). By definition, we have that for any $u \in \Delta(\mathcal{A})$ and $\vartheta \in \Delta(\mathcal{B})$

$$V(u, \vartheta_1) \leq V(u_1, \vartheta_1) \leq V(u_1, \vartheta), \qquad V(u, \vartheta_2) - \epsilon \leq V(u_2, \vartheta_2) \leq V(u_2, \vartheta) + \epsilon.$$

Then, we have

$$V(u_1, \vartheta_1) \geq V(u_2, \vartheta_1) \geq V(u_2, \vartheta_2) - \epsilon, \qquad V(u_1, \vartheta_1) \leq V(u_1, \vartheta_2) \leq V(u_2, \vartheta_2) + \epsilon.$$

Combining the two, we have

$$\begin{aligned} V(u, \vartheta_2) - 2\epsilon \leq V(u_2, \vartheta_2) - \epsilon \leq V(u_1, \vartheta_1) &\leq V(u_1, \vartheta_2) \\ &\leq V(u_2, \vartheta_2) + \epsilon \leq V(u_1, \vartheta_1) + 2\epsilon \leq V(u_1, \vartheta) + 2\epsilon \end{aligned}$$

for any $u \in \Delta(\mathcal{A})$ and $\vartheta \in \Delta(\mathcal{B})$, showing that (u_1, ϑ_2) is an 2ϵ-NE. The proof for the pair (u_2, ϑ_1) is analogous. ■