Minimax Sample Complexity for Turn-based Stochastic Game (Supplementary material)

Qiwen Cui[1] **Lin F. Yang**[2]

[1]School of Mathematical Sciences, Peking University
[2]Electrical and Computer Engineering Department, University of California, Los Angeles

Abstract

The empirical success of multi-agent reinforcement learning is encouraging, while few theoretical guarantees have been revealed. In this work, we prove that the plug-in solver approach, probably the most natural reinforcement learning algorithm, achieves minimax sample complexity for turn-based stochastic game (TBSG). Specifically, we perform planning in an empirical TBSG by utilizing a 'simulator' that allows sampling from arbitrary state-action pair. We show that the empirical Nash equilibrium strategy is an approximate Nash equilibrium strategy in the true TBSG and give both problem-dependent and problem-independent bound. We develop reward perturbation techniques to tackle the non-stationarity in the game and Taylor-expansion-type analysis to improve the dependence on approximation error. With these novel techniques, we prove the minimax sample complexity of turn-based stochastic game.

1 INTRODUCTION

Reinforcement learning (RL) [Sutton and Barto, 2018], a framework where agents learn to make sequential decisions in an unknown environment, has received tremendous attention. An interesting branch is multi-agent reinforcement learning (MARL) that multiple agents exist and they interact with the environment as well as the others, which bridges RL and game theory. In general, each agent attempts to maximize its own reward by utilizing the data collected from the environment and also inferring other agents' strategies. Impressive successes have been achieved in games such as backgammon [Tesauro, 1995], Go [Silver et al., 2017] and strategy games [Ye et al., 2020]. MARL has shown the potential for superhuman performance, but theoretical guarantees are rather rare due to complex interaction between agents that makes the problem considerably harder than single agent reinforcement learning. This is also known as non-stationarity in MARL, which means when multiple agents alter their strategies based on samples collected from previous strategy, the system becomes non-stationary for each agent and the improvement can not be guaranteed. One fundamental question in MBRL is that how to design efficient algorithms to overcome non-stationarity.

Two-players turn-based stochastic game (TBSG) is a two-agents generalization of Markov decision process (MDP), where two agents choose actions in turn and one agent wants to maximize the total reward while the other wants to minimize it. As a zero-sum game, TBSG is known to have Nash equilibrium strategy [Shapley, 1953], which means there exists a strategy pair that both agents will not benefit from changing its strategy alone, and our target is to find the (approximate) Nash equilibrium strategy. Dynamic programming type algorithms is a basic but powerful approach to solve TBSG. Strategy iteration, a counterpart of policy iteration in MDP, is known to be a polynomial complexity algorithm to solve TBSG with known transition kernel [Hansen et al., 2013, Jia et al., 2020]. However, these algorithms suffer from high computational cost and require full knowledge of the transition dynamic. Reinforcement learning is a promising alternative, which has demonstrated its potential in solving sequential decision making problems. However, non-stationarity pose a large obstacle against the convergence of model-free algorithms. To tackle this challenge, sophisticated algorithms have been proposed [Jia et al., 2019, Sidford et al., 2020]. In this work, we focus on another promising but insufficiently developed method, model-based algorithms, where agents learn the model of the environment and plan in the empirical model.

Model-based RL is long perceived to be the cure to the sample inefficiency in RL, which is also justified by empirical advances [Kaiser et al., 2019, Wang et al., 2019]. However, the theoretical understanding of model-based RL is still far from complete. Recently, a line of research focuses on

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

analyzing model-based algorithm under generative model setting [Azar et al., 2013, Agarwal et al., 2019, Cui and Yang, 2020, Zhang et al., 2020, Li et al., 2020], In this work, we aim to prove that the simplest model-based algorithm, plug-in solver approach, enjoys minimax sample complexity for TBSG by utilizing novel absorbing TBSG and reward perturbation techniques.

Specifically, we assume that we have access to a generative model oracle [Kakade et al., 2003], which allows sampling from arbitrary state-action pair. It is known that exploration and exploitation tradeoff is simplified under this setting, i.e., sampling from each state-action pair equally can already yield the minimax sample complexity result. With the generative model, the most intuitive approach is to learn an empirical TBSG and then use a planning algorithm to find the empirical Nash equilibrium strategy as the solution, which separates learning and planning. This kind of algorithm is known as the 'plug-in solver approach', which means *arbitrary* planning algorithm can be used. We will show that this simple plug-in solver approach enjoys minimax sample complexity. For MDP, this line of research is studied in the seminal work by Azar et al. [2013]. In the recent work [Li et al., 2020], they almost completely solves this problem by proving the minimax complexity with full range of error ϵ. However, the result for TBSG is still unknown and largely due to interaction between agents. In this work, we give the sample complexity upper bounds of TBSG for both problem-dependent and problem-independent cases.

Suboptimality gap is a widely studied notion in the bandit theory, which stands for a constant gap between the optimal arm and second optimal arm. This notion has received increasing focus in MDP and we generalize it to TBSG. To start with, we prove that $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\Delta^{-2})$ samples are enough to recover Nash equilibrium strategy accurately for $\Delta \in (0, (1-\gamma)^{-\frac{1}{2}}]$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, γ is the discount factor and Δ is the suboptimality gap. The key analysis tool is the absorbing TBSG technique that helps to show the empirical optimal Q value is close to true optimal Q value. With the absorbing TBSG, the statistical dependence on $\widehat{P}(s,a)$ is moved to a single parameter u, which can be approximated by an covering argument. Therefore, standard concentration arguments combined with union bound can be applied. Suboptimality gap plays an critical role in showing the empirical Nash equilibrium strategy is exactly the same as true Nash equilibrium strategy.

The main contribution is in the second part, where we give our problem-independent bound which meets the existing lower bound $O(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\epsilon^{-2})$ [Azar et al., 2013, Sidford et al., 2020]. Note that the problem-dependent result becomes meaningless when the suboptimality gap is sufficiently small and also the gap may not even exist. In the problem-independence case, we develop the reward perturbation technique to create such a gap in the estimated TBSG, which is inspired by the technique developed in [Li et al., 2020]. The key observation is that if $r(s,a)$ increases, $Q^*(s,a)$ increases much faster than $Q^*(s,a')$ for $a' \neq a$, thus the perturbed TBSG enjoys a suboptimality gap with high probability. Different from the usage in the first part, here the suboptimality gap is used to ensure the empirical Nash equilibrium strategy lies in a finite set so that union bound can be applied. Combining the reward perturbation technique with the absorbing TBSG technique, we are able to prove more subtle concentration arguments and finally show that the empirical Nash equilibrium strategy is an ϵ-approximate Nash equilibrium strategy in the true TBSG with minimax sample complexity $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\epsilon^{-2})$ for $\epsilon \in (0, (1-\gamma)^{-1}]$. In addition, we can recover the problem-dependent bound by simply setting $\epsilon = \Delta$.

Recently Zhang et al. [2020] proposes a similar result for simultaneous stochastic game. However, their result requires a planning oracle for regularized stochastic game, which is computationally intractable. Our algorithm only requires planning in a standard turn-based stochastic game, which can be performed efficiently by utilizing strategy iteration [Hansen et al., 2013] or learning algorithms [Sidford et al., 2020]. In addition, their sample complexity result $N = \widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\epsilon^{-2})$ only holds for $\epsilon \in (0, (1-\gamma)^{-\frac{1}{2}}]$, which means $N = \widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-2})$. our result fills the blank in the sample region $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-1}) \leq N \leq \widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-2})$. Note that the lower bound [Azar et al., 2013] indicates that $N = O(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-1})$ is insufficient to learn a policy. As both parts of our analysis heavily rely on the suboptimality gap, we hope our work can provide more understanding about this notion and TBSG.

2 PRELIMINARY

Turn-based Stochastic Game Turn-based two-player zero-sum stochastic game (TBSG) is a generalized version of Markov decison process (MDP) which includes two players competing with each other. Player 1 aims to maximize the total reward while player 2 aims to minimize it. TBSG is described by the tuple $\mathcal{G} = (\mathcal{S} = \mathcal{S}_{\max} \cup \mathcal{S}_{\min}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}_{\max}$ is the state space of player 1, $\mathcal{S}_{\min}$ is the state space of player 2, $\mathcal{A}$ is the action space of both players, $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ is the transition probability matrix, $r \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the reward vector and γ is the discount factor. In each step, only one player plays an action and a transition happens. For instance, if the state $s \in \mathcal{S}_{\max}$, player 1 needs to select an action $a \in \mathcal{A}$. After selecting the action, the state will transit to $s' \in \mathcal{S}_{\min}$ according to the distribution $P(\cdot|s,a)$ with reward $r(s,a)$ and player 2 needs to choose the action. For representation simplicity and without loss of generality, we assume that r is known and only P is

unknown[1].

We denote a strategy pair as $\pi := (\mu, \nu)$, where μ is the strategy of player 1 and ν is the strategy of player 2. Given strategy π, the value function and Q-function can be defined similarly as in MDP:

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s^t, \pi(s^t)) \,\middle|\, s^0 = s\right],$$

$$\begin{aligned} &Q^{\pi}(s,a) \\ :=&\mathbb{E}\left[r(s^0,a^0) + \sum_{t=1}^{\infty} \gamma^t r(s^t, \pi(s^t)) \,\middle|\, s^0 = s, a^0 = a\right] \\ =&r(s,a) + \gamma P(s,a) V^{\pi}. \end{aligned}$$

From the perspective of player 1, if the strategy ν of player 2 is given, TBSG degenerates to an MDP, so the optimal policy against ν exists, which we called as counterstrategy and use $c_{\max}(\nu)$ to denote it. Similarly we can define $c_{\min}(\mu)$ as the counterstrategy of μ for player 2. For simplicity, we ignore the subscript in $c_{\max}$ and $c_{\min}$ when it is clear in the context. In addition, we define $V^{*,\nu} := V^{c(\nu),\nu}$ and $V^{\mu,*} := V^{\mu,c(\mu)}$ and the same for Q. By definition and property of optimal policy in MDP, we have

$$\begin{aligned} Q^{*,\nu}(s,a) &= \max_{\mu} Q^{\mu,\nu}(s,a), \forall s \in \mathcal{S}, \\ Q^{\mu,*}(s,a) &= \min_{\nu} Q^{\mu,\nu}(s,a), \forall s \in \mathcal{S}, \\ Q^{*,\nu}(s, c_{\max}(\nu)(a)) &= \max_{a'} Q^{*,\nu}(s,a'), \forall s \in \mathcal{S}_{\max}, \\ Q^{\mu,*}(s, c_{\min}(\mu)(s)) &= \min_{a'} Q^{\mu,*}(s,a'), \forall s \in \mathcal{S}_{\min}. \end{aligned}$$

Note that these are the sufficient and necessary condition of counterstrategy, which will be utilized repeatedly in our analysis.

To solve a TBSG, our goal is to find the Nash equilibrium strategy $\pi^* = (\mu^*, \nu^*)$, where $\mu^* = c(\nu^*), \nu^* = c(\mu^*)$. For Nash equilibrium strategy, neither player can benefit from changing its policy alone. As μ^* and ν^* are counterstrategy to each other, they inherit properties of counterstrategy given above. For simplicity, we will not repeat here. It is well known that in TBSG, there always exists a pure strategy as the Nash equilibrium strategy. In addition, all Nash equilibrium strategy share the same state-action value, which makes pure Nash equilibrium strategy unique given some tie selection rule, so we only consider pure strategies in our analysis.

Specifically, our target is to find an ϵ-approximate Nash equilibrium strategy $\pi = (\mu, \nu)$ such that

$$\left|Q^{\mu,*}(s,a) - Q^*(s,a)\right| \le \epsilon, \forall (s,a),$$

$$\left|Q^{*,\nu}(s,a) - Q^*(s,a)\right| \le \epsilon, \forall (s,a),$$

for some $\epsilon > 0$ with as few samples as possible. Note that this is different from and stronger than the MDP analogue, which should be $\left|Q^{\pi}(s,a) - Q^*(s,a)\right| \le \epsilon$. This slight difference makes subtle difficulty as we will show later.

Generative Model Oracle and Plug-in Solver Approach We assume that we have access to a generative model, where we can input an arbitrary state action pair (s,a) and receive a sampled state form $P(\cdot|s,a)$. Generative model oracle was introduced in [Kearns and Singh, 1999, Kakade et al., 2003]. This setting is different from the offline oracle where we can only sample trajectories via a behaviour policy and online oracle where we adaptively change the policy to explore and exploit. The advantage of generative model setting is that the exploration and exploitation is simplified, as previous work shows that treating all state-action pair equally is already optimal [Azar et al., 2013, Sidford et al., 2018]. In particular, we call the generative model $N/|\mathcal{S}||\mathcal{A}|$ times on each state-action pair and construct the empirical TBSG $\widehat{G} = (\mathcal{S} = \mathcal{S}_{\max} \cup \mathcal{S}_{\min}, \mathcal{A}, \widehat{P}, r, \gamma)$:

$$\widehat{P}(s'|s,a) = \frac{\text{count}(s,a,s')}{N/|\mathcal{S}||\mathcal{A}|}, \forall s, s' \in \mathcal{S}_{\max} \cup \mathcal{S}_{\min}, a \in \mathcal{A},$$

where $\text{count}(s,a,s')$ is the number of times that s' is sampled from state-action pair (s,a). It is straightforward that $\widehat{P}$ is an unbiased and maximum likelihood estimation of the true transition kernel P. We use $\widehat{\pi}^* = (\widehat{\mu}^*, \widehat{\nu}^*)$ to denote the Nash equilibrium strategy in the empirical MDP as well as $\widehat{V}$ and $\widehat{Q}$. The algorithm is given in Algorithm 1.

As the transition kernel in $\widehat{\mathcal{G}}$ is known, arbitrary planning algorithm can be used to find the empirical Nash equilibrium strategy $\widehat{\pi}^*$. One choice is to use strategy iteration, which finds $\widehat{\pi}^*$ with in $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-1})$ iterations [Hansen et al., 2013]. In addition, algorithms that find approximate Nash equilibrium strategy can be applied, such as QVI-MIVSS in [Sidford et al., 2020]. Note that our analysis is for the exact Nash equilibrium strategy $\widehat{\pi}^*$, but it can be generalized to approximate Nash equilibrium strategy by using techniques in [Agarwal et al., 2019].

Suboptimality Gap Suboptimality gap is originated in bandit theory, which is the gap between the mean reward of the best arm and second best arm. In TBSG, we define the suboptimality gap based on optimal Q-value and second optimal Q-value.

Definition 1. (Suboptimality gap for Nash equilibrium strategy) A TBSG enjoys a suboptimality gap of Δ for Nash equilibrium strategy if and only if

$$\forall s \in \mathcal{S}_{\max}, a \neq \mu^*(s) : Q^*(s, \mu^*(s)) - Q^*(s,a) \ge \Delta,$$

$$\forall s \in \mathcal{S}_{\min}, a \neq \nu^*(s) : Q^*(s, \nu^*(s)) - Q^*(s,a) \le -\Delta,$$

and if there are two optimal actions, the gap is zero.

[1] Our proof can be easily adapted to show that the sample complexity of learning the reward r is an order of $\frac{1}{1-\gamma}$ smaller than learning the transition P.

Definition 2. (Suboptimality gap for counterstrategy) A TBSG enjoys a suboptimality gap of Δ for counter strategy to the strategy ν of player 2 if and only if

$$\forall s \in \mathcal{S}_{\max}, a \neq c(\nu)(s), Q^*(s, c(\nu)(s)) - Q^*(s, a) \geq \Delta.$$

A TBSG enjoys a suboptimality gap of Δ for counter strategy to the strategy μ of player 1 if and only if

$$\forall s \in \mathcal{S}_{\min}, a \neq c(\mu)(s), Q^*(s, c(\mu)(s)) - Q^*(s, a) \leq -\Delta.$$

The suboptimality gap means that following the Nash equilibrium strategy, the expected total reward of the best action and the second best action differs at least Δ. Intuitively, this gap quantifies the difficulty of learning the optimal action and a small gap hinders finding out the optimal action.

Notations $f(x) = O(g(x))$ means that there exists a constant C such that $f \leq Cg$ and $f(x) = \Omega(g(x))$ means that $g(x) = O(f(x))$. $\widetilde{O}$ and $\widetilde{\Omega}$ is same as O and Ω except that logarithmic factors are ignored. We use $f \gtrsim g$ to denote that there exist some constant C such that $f \geq Cg$ and $f \lesssim g$ means $f \leq Cg$ for some constant C. For a strategy π, $P^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ is the transition matrix induced by policy π and $P^\pi(s,a)(s',a') = P(s'|s,a)\mathbf{1}(a' = \pi(s'))$ where $\mathbf{1}(\cdot)$ is the indicator function. $P(s,a)$ is the row vector of P that correspond to s, a. We use $|\cdot|$ to denote the infinity norm $\|\cdot\|_\infty$ and $\sqrt{\cdot}, \leq, \geq$ are entry-wise operators.

Algorithm 1: Solving TBSG via Plug-in Solver

Input: A generative model that can output samples from distribution $P(\cdot|s,a)$ for query (s,a), a plug-in solver.

Initial: Sample size: N;

for *(s,a) in* $\mathcal{S} \times \mathcal{A}$ **do**

  Collect $N/|\mathcal{S}||\mathcal{A}|$ samples from $P(\cdot|s,a)$;

  Compute $\widehat{P}(s'|s,a) = \frac{count(s,a,s')}{N/|\mathcal{S}||\mathcal{A}|}$;

end

Construct the (perturbed) empirical TBSG;

$\widehat{\mathcal{G}} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma)$;

$\widehat{\mathcal{G}}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, \widehat{P}, r_{\mathrm{p}}, \gamma)$;

Plan with an arbitrary plug-in solver and receive the (perturbed) empirical Nash equilibrium strategy $\widehat{\pi}^*$ ($\widehat{\pi}^*_{\mathrm{p}}$);

Output: $\widehat{\pi}^*$ ($\widehat{\pi}^*_{\mathrm{p}}$)

3 TECHNICAL LEMMAS FROM MDP

In this section, we present several technical lemmas that is originated in MDP analysis [Azar et al., 2013, Agarwal et al., 2019, Li et al., 2020]. These lemmas can be easily adapted to TBSG and we present them to give some intuition on our TBSG analysis.

Lemma 1. *For any strategy π, we have*

$$\begin{aligned} Q^\pi - \widehat{Q}^\pi &= \gamma(I - P^\pi)^{-1}(\widehat{P} - P)\widehat{V}^\pi \\ &= \gamma(I - \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi. \end{aligned}$$

This lemma portrays the concentration of $\widehat{Q}^\pi$. Note that there are two kinds of factorization and requires different analysis. In the problem-dependent bound, we use the first one and in the problem-independent bound, we use the second one.

Definition 3. (One-step Variance) We define the one-step variance of a state-action pair (s,a) with respect to a certain value function V to be

$$Var_{s,a}(V) := P(s,a)V^2 - (P(s,a)V)^2,$$

which is the variance of the next state value. We define $Var_P(V) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to be a vector consisting of all $Var_{s,a}(V)$.

We define the one-step variance to facilitate the usage of Bernstein's inequality on term $(\widehat{P} - P)\widehat{V}^\pi$. A detailed introduction of variance in MDP can be found in Azar et al. [2013]. The following two lemmas show how to bound the one-step variance term.

Lemma 2. *For any policy π and V^π is the value function in a MDP with transition P, we have*

$$\left|(I - \gamma P^\pi)^{-1}\sqrt{Var_P(\widehat{V}^\pi)}\right| \leq \sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\left|Q^\pi - \widehat{Q}^\pi\right|}{1-\gamma}.$$

Lemma 3. *For any policy π and V^π is the value function in a MDP with transition P, if $N \gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}\log\left(\frac{1}{(1-\gamma)\delta}\right)$, with probability larger than $1-\delta$, we have*

$$\left|(I - \gamma\widehat{P}^\pi)^{-1}\sqrt{Var_P(V^\pi)}\right| \leq \frac{16}{\sqrt{(1-\gamma)^3}}.$$

Lemma 2 correspond to the first factorization in Lemma 1 and Lemma 3 is correspond to the second one. Lemma 2 is derived by utilizing the variance-Bellman-equation and Lemma 3 is derived by a taylor expansion type analysis. If we can apply Bernstein's inequality to $(\widehat{P} - P)\widehat{V}^\pi$ or $(\widehat{P} - P)V^\pi$ to generate the $\sqrt{Var_P(\widehat{V}^\pi)}$ or $\sqrt{Var_P(V^\pi)}$ term, then with Lemma 2 or Lemma 3, we can bound $|Q^\pi - \widehat{Q}^\pi|$.

By concentration inequalities, we can bound $(P - \widehat{P})V^\pi$ for *fixed* π. However, if $\pi = (\widehat{\mu}^*, c(\widehat{\mu}^*))$ or $\pi = (\mu^*, \widehat{c}(\mu))$, the complex statistical dependence between π and $\widehat{P}$ hinders the conventional concentration and subtle techniques are needed. In addition, $(P - \widehat{P})\widehat{V}^\pi$ suffers from the dependence even for fixed π. The key contribution in the next two section is to use novel TBSG techniques to make the concentration arguments applicable to these two terms.

4 WARM UP: PROBLEM-DEPENDENT UPPER BOUND

In this section, we show that if a TBSG $\mathcal{G}$ enjoys a suboptimality gap of Δ, then with $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\Delta^{-2})$ samples, the empirical Nash equilibrium strategy $\widehat{\pi}^*$ in $\widehat{\mathcal{G}}$ is exactly the Nash equilibrium strategy π^* in $\mathcal{G}$ with high probability. To begin with, we introduce a novel absorbing TBSG technique, which is motivated by the absorbing MDP technique developed in [Agarwal et al., 2019]. An absorbing TBSG $\widetilde{\mathcal{G}}_{s,a,u}$ is identical to the empirical TBSG $\widehat{\mathcal{G}}$ except that the transition distribution of a specific state-action pair (s,a) is set to be absorbing. Similar techniques have been developed for MDP [Li et al., 2020] and simultaneous game [Zhang et al., 2020].

Definition 4. (Absorbing TBSG) For a TBSG $\mathcal{G} = (\mathcal{S}_{max}, \mathcal{S}_{min}, \mathcal{A}, P, r, \gamma)$ and a given state-action pair $(s,a) \in (\mathcal{S}_{max} \cup \mathcal{S}_{min}) \times \mathcal{A}$, the absorbing TBSG $\widetilde{\mathcal{G}}_{s,a,u} = (\mathcal{S}_{max}, \mathcal{S}_{min}, \mathcal{A}, \widetilde{P}, \widetilde{r}, \gamma)$, where

$$\widetilde{P}(s|s,a) = 1, \widetilde{P}(\cdot|s',a') = P(\cdot|s',a'), \forall (s',a') \neq (s,a),$$

$$\widetilde{r}(s,a) = u, \widetilde{r}(s',a') = r(s',a'), \forall (s',a') \neq (s,a).$$

Remark 1. *Absorbing TBSG is independent of $\widehat{P}(s,a)$, which is a kind of leave-one-out analysis. Note that $\widetilde{P}(s,a)$ can be set to arbitrary fixed distribution. We use the absorbing distribution for simplicity and correspondence to its name.*

For simplicity, we ignore (s,a) in absorbing TBSG when there is no misunderstanding. We use $\widetilde{\pi}_u, \widetilde{Q}_u, \widetilde{V}_u$ to denote the strategy, state-action value and state value in the absorbing TBSG. These terms actually depend on (s,a) and we omit (s,a) when there is no confusion. Note that $\widetilde{\mathcal{G}}_u$ has no dependence on $\widehat{P}(s,a)$, which makes the concentration on $(\widehat{P}(s,a) - P(s,a))\widetilde{V}_u$ possible. The key property of absorbing TBSG is that it can recover the Q-value of the empirical MDP by tuning the parameter u. Moreover, the Q-value of absorbing TBSG is $\frac{1}{1-\gamma}$-lipschitz to u, which means we can use an ϵ-cover on the range of u to approximately recover the Q value in the empirical TBSG.

Lemma 4. *(Properties of absorbing TBSG) Set $u^* = r(s,a) + \gamma(P(s,a)V^*) - \gamma V^*(s), u^\mu = r(s,a) + \gamma(P(s,a)V^\mu) - \gamma V^\mu(s)$. Then we have*

$$Q^*_{u^*} = Q^*, Q^{\mu,*}_{u^\mu} = Q^{\mu,*},$$

$$|Q^*_u - Q^*_{u'}| \leq \frac{|u-u'|}{1-\gamma}, |Q^{\mu,*}_u - Q^{\mu,*}_{u'}| \leq \frac{|u-u'|}{1-\gamma}.$$

We can concentrate $(\widehat{P}(s,a) - P(s,a))\widehat{V}^{\mu,*}$ by a union bound and an additional approximation error term as we construct an absorbing TBSG $\widetilde{\mathcal{G}}$ on empirical TBSG $\widehat{\mathcal{G}}$. Combining Lemma 1 and Lemma 2, we can bound $|Q^{\mu,\widehat{c}(\mu)} - \widehat{Q}^{\mu,c(\mu)}|$.

Lemma 5.

$$\left|Q^* - \widehat{Q}^*\right| \leq \max\{\left|Q^{\mu,\widehat{c}(\mu)} - \widehat{Q}^{\mu,*}\right|, \left|Q^{\widehat{c}(\nu),\nu} - \widehat{Q}^{*,\nu}\right|\}.$$

Lemma 5 shows that we can bound the error of estimate the optimal state-action value $\left|Q^* - \widehat{Q}^*\right|$ by policy evaluation error $\left|Q^{\mu,\widehat{c}(\mu)} - \widehat{Q}^{\mu,*}\right|$ and $\left|Q^{\widehat{c}(\nu),\nu} - \widehat{Q}^{*,\nu}\right|$.

Now we show how to use absorbing TBSG and suboptimality gap to prove that $\widehat{\pi}^* = \pi^*$ with large probability. The key is to prove that if $|Q^* - \widehat{Q}^*| \leq \frac{\Delta}{2}$, then by the definition of suboptimality gap, $\widehat{Q}^*$ enjoys the gap for policy π^*. As $\widehat{\pi}^*$ is the Nash equilibrium policy in $\widehat{\mathcal{G}}$, it is the only policy that can enjoy the gap, which means $\widehat{\pi}^* = \pi^*$.

Lemma 6. *If Q^* enjoys a suboptimality gap of Δ and $\left|Q^* - \widehat{Q}^*\right| \leq \frac{\Delta}{2}$, then we have $\widehat{\pi}^* = \pi^*$.*

Combining all the parts together, we get our problem-dependent result. Theorem 1 indicates that with a suboptimality gap, we can accurately recover the Nash equilibrium strategy with polynomial sample complexity. The detailed proof is provided in Appendix B.

Theorem 1. *If $\mathcal{G}$ enjoys a suboptimality gap of Δ and the number of samples satisfies*

$$N \geq \frac{C|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\Delta^2} \log(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\Delta})$$

for some constant C and $\Delta \in (0, (1-\gamma)^{-\frac{1}{2}}]$, then with probability at least $1-\delta$, we have $\widehat{\pi}^* = \pi^*$, which means the empirical Nash equilibrium strategy we obtained is exactly the Nash equilibrium strategy in the true TBSG.

Note that this result can recover π^* exactly, but at a price of restrictive sample complexity. We need to know the suboptimality gap Δ beforehand and we have little knowledge about the empirical Nash equilibrium strategy if the sample complexity is smaller than $\frac{C|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\Delta^2} \log(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\Delta})$. In the next section, we will give a problem-independent result, which has no dependence on the suboptimality gap of $\mathcal{G}$.

5 PROBLEM-INDEPENDENT UPPER BOUND

Two flaws exist in our problem-dependent result. One is if the suboptimality gap is considerably small, the bound $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\Delta^{-2})$ becomes meaningless. In addition, when $N < |\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\Delta^{-2}$, the quality of empirical

Nash equilibrium strategy $\widehat{\pi}^*$ is completely unknown. In this section, we aim to give a minimax sample complexity result without the assumption of suboptimality gap. Interestingly, though the suboptimality gap is not assumed, we artificially create the suboptimality gap instead and the role of suboptimality gap is different from the first part analysis, which we will specify later. We now introduce the reward perturbation technique that can artificially create a suboptimality gap.

5.1 REWARD PERTURBATION TECHNIQUE

Here we use a reward perturbation technique to create a suboptimality gap in TBSG, which is inspired by a similar argument in MDP analysis [Li et al., 2020]. We give a proof that is different from the one in [Li et al., 2020] and our analysis for TBSG can generalize to MDP automatically as MDP is a degenerated version of TBSG. We show that by randomly perturb the reward function, with large probability, the perturbed TBSG enjoys a suboptimality gap. First we define the perturbed TBSG.

Definition 5. (Perturbed TBSG) For a TBSG $\mathcal{G} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, the perturbed TBSG is $\mathcal{G}_{\mathrm{p}} = (\mathcal{S}, \mathcal{A}, P, r_{\mathrm{p}}, \gamma)$, where

$$r_{\mathrm{p}} = r + \zeta$$

and $\zeta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is a vector composed of independent random variables following uniform distribution on $[0, \xi]$.

We use subscript π_{p} to denote the strategy in perturbed TBSG as well as V_{p} and Q_{p}. The key property of perturbed TBSG is that it enjoy a suboptimality gap of $\frac{\xi\delta(1-\gamma)}{4|\mathcal{S}|^2|\mathcal{A}|}$ with probability at least $1-\delta$.

Lemma 7. *(TBSG version of suboptimality gap lemma in [Li et al., 2020]) For a fixed policy ν, with probability at least $1-\delta$, the perturbed TBSG $\mathcal{G}_{\mathrm{p}}$ enjoys a suboptimality gap of $\frac{\xi\delta(1-\gamma)}{4|\mathcal{S}|^2|\mathcal{A}|}$ for the Nash equilibrium strategy and a same gap for counterstrategy of ν.*

Our proof is substantially different from [Li et al., 2020], which consists of two important observation. Here we consider the case where only $r(s, a_1)$ is perturbed to $r(s, a_1) + \tau$. First, the Nash equilibrium strategy π^*_τ is a piecewise constant function of τ, if some tie breaking rule is given. Second, we show that $Q^*_\tau(s, a_1) = k_1(\pi_\tau)\tau + b_1(\pi_\tau)$ and $Q^*_\tau(s, a_2) = k_2(\pi_\tau)\tau + b_2(\pi_\tau)$ are piecewise linear function and $k_2(\pi_\tau) \le \gamma k_1(\pi_\tau), \forall \tau$. Intuitively, these two observation means that $Q^*_\tau(s, a_2)$ grows at most γ times the speed of $Q^*_\tau(s, a_1)$, which further implies $|Q^*_\tau(s, a_1) - Q^*_\tau(s, a_2)| \le w$ can only holds for a small interval of τ.

Lemma 8. *Consider a TBSG $\mathcal{G}_\tau = (\mathcal{S}, \mathcal{A}, P, r_\tau, \gamma)$ where $r_\tau = r + \tau \mathbf{1}_{s,a}$. Then the following facts hold.*

- *Given a rule to select optimal action when there are multiple optimal actions, π^*_τ is a constant (vector) function of τ.*
- *Q^*_τ is a piecewise linear (vector) function of τ.*
- *$Q^*_\tau(s, a') = kQ^*_\tau(s, a) + b$, where $0 \le k \le \gamma$ and b are a function of π^*_τ.*

With the above argument, we can prove that for arbitrary s, a, a', if we increase $r(s, a)$, then the growth of $Q^*_\tau(s, a')$ is at most γ times of $Q^*_\tau(s, a)$, which means $|Q^*_\tau(s, a') - Q^*_\tau(s, a)|$ only holds for a small range of $r_\tau(s, a)$. With a union bound argument, we prove the existence of suboptimality gap. The proof for counterstrategy is similar and the details are given in the appendix.

5.2 MINIMAX SAMPLE COMPLEXITY

In this section, we show how to use the reward perturbation technique to derive the minimax sample complexity result. First, we show that the optimal strategy in perturbed empirical TBSG is contained in a finite set that has no dependence on $\widehat{P}(s, a)$.

Lemma 9. *Set U to be a set of equally spaced points in $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$ and $|U| = \frac{16|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^2\xi\delta}$. We define*

$$\mathcal{M}^* = \{\widetilde{\mu}^*_{\mathrm{p},u} : u \in U\}, \mathcal{M}^{\mu^*} = \{\widetilde{c}_{\mathrm{p},u}(\mu^*) : u \in U\}.$$

With probability at least $1-\delta$, we have $\widehat{\mu}^*_{\mathrm{p}} \in \mathcal{M}^*$ and $\widehat{c}_{\mathrm{p}}(\mu^*) \in \mathcal{M}^{\mu^*}$.

Lemma 9 means with large probability $\widehat{\mu}^*_{\mathrm{p}}$ and $\widehat{c}_{\mathrm{p}}(\mu^*)$ lie in a finite set, which is independent of $\widehat{P}(s, a)$. This independence allows the usage of Bernstein's inequality and with the union bound, we can prove the concentration of $(\widehat{P}(s, a) - P(s, a))V_{\mathrm{p}}^{\widehat{\mu}^*_{\mathrm{p}},*}$ and $(\widehat{P}(s, a) - P(s, a))V_{\mathrm{p}}^{\mu^*,\widehat{c}_{\mathrm{p}}(\mu^*)}$. Then, with Lemma 1 and Lemma 3, we can bound $|Q_{\mathrm{p}}^{\widehat{\mu}^*_{\mathrm{p}},*} - \widehat{Q}_{\mathrm{p}}^{\widehat{\mu}^*_{\mathrm{p}},c(\widehat{\mu}^*_{\mathrm{p}})}|$ and $|Q_{\mathrm{p}}^{\mu^*,\widehat{c}_{\mathrm{p}}(\mu^*)} - \widehat{Q}_{\mathrm{p}}^{\mu^*,*}|$.

Lemma 10. *For perturbed empirical Nash equilibrium strategy $\widehat{\pi}^*_{\mathrm{p}} = (\widehat{\mu}^*_{\mathrm{p}}, \widehat{\nu}^*_{\mathrm{p}})$, we have*

$$\begin{aligned}&|Q^{\widehat{\mu}_{\mathrm{p}},*} - Q^*| \\ \le&|Q_{\mathrm{p}}^{\widehat{\mu}^*_{\mathrm{p}},*} - \widehat{Q}_{\mathrm{p}}^{\widehat{\mu}^*_{\mathrm{p}},c(\widehat{\mu}^*_{\mathrm{p}})}| + |Q_{\mathrm{p}}^{\mu^*,\widehat{c}_{\mathrm{p}}(\mu^*)} - \widehat{Q}_{\mathrm{p}}^{\mu^*,*}| + \frac{4\xi}{1-\gamma}.\end{aligned}$$

Lemma 10 shows how to bound $|Q^{\widehat{\mu}_{\mathrm{p}},*} - Q^*|$ by perturbed TBSG. In the same manner, we can bound $|Q^{*,\widehat{\nu}_{\mathrm{p}}} - Q^*|$. Selecting an appropriate ξ, we can show that perturbed empirical Nash equilibrium strategy $\widehat{\pi}^*_{\mathrm{p}}$ is an ϵ-Nash equilibrium strategy. The detailed proof is provided in Appendix C.

Theorem 2. *If the number of samples satisfies*

$$N \geq \frac{C|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\log(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\epsilon})$$

for some constant C, then with probability at least $1-\delta$ and $\epsilon \in (0,(1-\gamma)^{-1}]$, we have that $\widehat{\pi}^*_{\mathrm{p}}$ is an ϵ-approximate Nash equilibrium strategy in $\mathcal{G}$.

Remark 2. *Compared with Theorem 3.6 in [Zhang et al., 2020], Theorem 2 extends the range of epsilon to $(0,(1-\gamma)^{-1}]$, which is the full possible range. In addition, Zhang et al. [2020] requires a smooth planning oracle, which is computationally inefficient. Our algorithm only needs a standard planning oracle to solve the empirical TBSG.*

In addition, we can derive the following improved result for problem-dependent bound by choosing $\epsilon = \frac{\Delta}{2}$ and further analysis on suboptimality gap, which is provided in the Appendix C.

Theorem 3. *If $\mathcal{G}$ enjoys a suboptimality gap of Δ and the number of samples satisfies*

$$N \geq \frac{C|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\Delta^2}\log(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\Delta})$$

for some constant C and $\Delta \in (0,(1-\gamma)^{-1}]$, then with probability at least $1-\delta$, we have $\widehat{\pi}^*_{\mathrm{p}} = \pi^*$, which means the empirical Nash equilibrium strategy we obtained is exactly the Nash equilibrium strategy in the true TBSG.

Theorem 3 is strictly stronger than Theorem 1. The difference between Theorem 1 and Theorem 3 is that the range of suboptimality gap Δ is extended to $(0,(1-\gamma)^{-1}]$, which is the full possible range of suboptimality gap.

6 RELATED LITERATURE

TBSG TBSG has been widely studied since [Shapley, 1953]. For a detailed introduction of stochastic game, readers can refer to [Neyman et al., 2003]. In the old days, people focus on dynamic programming type algorithms to solve TBSG. Strategy iteration, as the counterpart of value iteration in MDP and parallelized simplex method, is proved to be a strong polynomial time algorithm [Hansen et al., 2013, Jia et al., 2020]. Reinforcement learning approach has been studied recently for TBSG to relieve the high computational cost of dynamic programming. Several works are proposed for the generative model setting, [Sidford et al., 2020, Jia et al., 2019, Zhang et al., 2020]. [Sidford et al., 2020] first gives a sample efficient algorithm for tabular TBSG, while their result achieves minimax sample complexity $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\epsilon^{-2})$ only for $\epsilon \in (0,1]$. [Jia et al., 2019] adapts the MDP algorithm in [Sidford et al., 2018] for feature-based TBSG, but leaves a gap of $\frac{1}{1-\gamma}$ between optimal sample complexity. [Cui and Yang, 2020] uses a similar algorithm as ours in feature-based TBSG, while their result only holds for ϵ-Nash equilibrium value, which results in a $\frac{1}{(1-\gamma)^2}$ gap in finding ϵ-Nash equilibrium strategy. [Zhang et al., 2020] considers simultaneous stochastic game, and their approach consists of solving a regularized simultaneous stochastic game, which is computationally costly. For the online sampling setting, a recent work [Bai et al., 2020] uses an upper confidence bound algorithms that can find an approximate Nash equilibrium strategy in $\widetilde{O}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|)$ steps.

Generative Model Generative model is a sampling oracle setting in MDP, which has been shown to simplify the exploration and exploitation tradeoff. This concept is formalized in [Kakade et al., 2003] and a $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\mathrm{poly}((1-\gamma)^{-1})\epsilon^{-2})$ sample complexity has been proved there. [Azar et al., 2013] proves the minimax sample complexity $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\epsilon^{-2})$. However, the upper bound there is only for $\epsilon \in (1-\gamma)^{-1/2}|\mathcal{S}|^{-1/2}$. Many works have devoted to improve the dependence on ϵ. Recently, [Sidford et al., 2018] gives a minimax model-free algorithm for $\epsilon \in (0,1]$ and [Agarwal et al., 2019] gives a minimax model-based algorithm for $\epsilon \in (0,(1-\gamma)^{-1/2}]$. Finally, [Li et al., 2020] uses a perturbed MDP technique to prove minimax sample complexity with full range of ϵ.

Suboptimality Gap Suboptimality gap originated in bandit theory. Multi-armed bandits and linear bandits enjoy a logarithmic gap-dependent regret and a square root gap-independent regret [Auer et al., 2002, Abbasi-Yadkori et al., 2011]. MDP with suboptimality gap have been studied in [Auer et al., 2009] and recently $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\mathrm{poly}(H)\log(T))$ regret has been proved for both model-based and model free algorithms [Simchowitz and Jamieson, 2019, Yang et al., 2020]. [Du et al., 2020] utilized the suboptimality gap in general function approximation setting and proved that optimal policy can be found in $\widetilde{O}(\dim_{\mathrm{E}})$ trajectories in deterministic MDP. Most of the gap-dependent analysis in MDP focus on online RL and to the best of our knowledge, we are the first to study this notion in TBSG with a generative model.

7 CONCLUSION

In this work, we completely solve the sample complexity problem of TBSG with generative model oracle. We prove that the simplest model-based algorithm, plug-in solver approach, is minimax sample optimal for full range of ϵ by using absorbing TBSG and reward perturbation techniques. Our proof is based on suboptimality gap, a notion originated from bandit theory and receives great attention in RL. We believe that our work can shed some light on suboptimality gap and TBSG.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Alekh Agarwal, Sham Kakade, and Lin F Yang. On the optimality of sparse model-based planning for markov decision processes. *arXiv preprint arXiv:1906.03804*, 2019.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

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.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Qiwen Cui and Lin F. Yang. Is plug-in solver sample-efficient for feature-based reinforcement learning?, 2020.

Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic q-learning with function approximation in deterministic systems: Tight bounds on approximation error and sample complexity. *arXiv preprint arXiv:2002.07125*, 2020.

Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

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

Zeyu Jia, Zaiwen Wen, and Yinyu Ye. Towards solving 2-tbsg efficiently. *Optimization Methods and Software*, 35(4):706–721, 2020.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

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

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.

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.

Abraham Neyman, Sylvain Sorin, and S Sorin. *Stochastic games and applications*, volume 570. Springer Science & Business Media, 2003.

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

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM, 2018.

Aaron Sidford, Mengdi Wang, Lin 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*, pages 2992–3002. PMLR, 2020.

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.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, pages 1153–1162, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019.

Kunhe Yang, Lin F Yang, and Simon S Du. q-learning with logarithmic regret. *arXiv preprint arXiv:2006.09118*, 2020.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *AAAI*, pages 6672–6679, 2020.

Kaiqing Zhang, Sham M Kakade, Tamer Başar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

A TECHNICAL LEMMAS FROM MDP

Additional Notations We use $\Pi^\pi : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|}$ to denote the projection operator with respect to policy π, which means if $V = \Pi^\pi Q$, then $V(s) = Q(s, \pi(s)), \forall s \in \mathcal{S}$. We use $\widehat{\mathcal{G}}$ to denote the empirical TBSG, $\widetilde{\mathcal{G}}$ to denote the absorbing TBSG on empirical TBSG and $\mathcal{G}_{\mathrm{p}}$ to denote the perturbed TBSG. We use $\mathbf{1}$ to denote a vector with all entries to be 1 and $\mathbf{1}_{s,a}$ to denote a zero vector with only (s, a) entry to be 1.

Proof of Lemma 1. By Bellman equation, we have $Q^\pi = (I - \gamma P^\pi)^{-1} r$ and $\widehat{Q}^\pi = (I - \gamma \widehat{P}^\pi)^{-1} r$. For any policy π, we have

$$\begin{aligned} Q^\pi - \widehat{Q}^\pi &= (I - \gamma P^\pi)^{-1} r - (I - \gamma \widehat{P}^\pi)^{-1} r \\ &= (I - \gamma P^\pi)^{-1} ((I - \gamma \widehat{P}^\pi) - (I - \gamma P^\pi)) \widehat{Q}^\pi \\ &= \gamma (I - \gamma P^\pi)^{-1} (P^\pi - \widehat{P}^\pi) \widehat{Q}^\pi \\ &= \gamma (I - \gamma P^\pi)^{-1} (P - \widehat{P}) \widehat{V}^\pi. \end{aligned}$$

Similarly, we have $Q^\pi - \widehat{Q}^\pi = \gamma (I - \gamma \widehat{P}^\pi)^{-1} (\widehat{P} - P) V^\pi$.

□

Lemma 11. *(Lemma 3 in [Sidford et al., 2020]) For any policy π, we have*

$$\left| (I - \gamma P^\pi)^{-1} \sqrt{\mathrm{Var}_{\mathrm{P}}(V^\pi)} \right| \leq \sqrt{\frac{2}{(1-\gamma)^3}}.$$

Proof of Lemma 2. We have

$$\begin{aligned} \left| (I - \gamma P^\pi)^{-1} \sqrt{\mathrm{Var}_{\mathrm{P}}(\widehat{V}^\pi)} \right| &\leq \left| (I - \gamma P^\pi)^{-1} \sqrt{\mathrm{Var}_{\mathrm{P}}(V^\pi)} \right| + \left| (I - \gamma P^\pi)^{-1} \sqrt{\mathrm{Var}_{\mathrm{P}}(\widehat{V}^\pi - V^\pi)} \right| \\ &\leq \sqrt{\frac{2}{(1-\gamma)^3}} + \frac{|\widehat{V}^\pi - V^\pi|}{1-\gamma} \\ &\leq \sqrt{\frac{2}{(1-\gamma)^3}} + \frac{|\widehat{Q}^\pi - Q^\pi|}{1-\gamma}, \end{aligned}$$

where the first inequality is the triangle inequality, the second one is from Lemma 11 and the fact that $|(I - \gamma P^\pi)^{-1} V| \leq \frac{|V|}{1-\gamma}$, and the last one is because V^π is a subset of Q^π.

□

Lemma 12. *(Lemma 8 in [Li et al., 2020]) Let Q be a vector obeying $Q = (I - \gamma P^\pi)^{-1} r$ for some vector $r > 0$ and $V = \Pi^\pi Q$, then we have*

$$\left| (I - \gamma P^\pi)^{-1} \sqrt{\mathrm{Var}_{\mathrm{P}}(V)} \right| \leq \frac{4}{\gamma \sqrt{1-\gamma}} |Q|.$$

Proof of Lemma 3. We use a Taylor expansion type analysis. For simplicity, we define

$$r^{(n)} = \sqrt{\mathrm{Var}_{\mathrm{P}}(V^{(n-1)})}, Q^{(n)} = (I - \gamma P^\pi)^{-1} r^{(n)}, V^{(n)} = \Pi^\pi Q^{(n)}, n = 1, 2, \cdots$$

and $V^{(0)} = V^\pi$. Then with large probability we have

$$\begin{aligned}
|(I-\gamma\widehat{P}^\pi)^{-1}\sqrt{\mathrm{Var}_{\mathrm{P}}(V^\pi)}| &= |(I-\gamma\widehat{P}^\pi)^{-1}r^{(1)}| \\
&= |(I-\gamma P^\pi)^{-1}r^{(1)} + (I-\gamma\widehat{P}^\pi)^{-1}(\gamma\widehat{P}^\pi - \gamma P^\pi)(I-\gamma P^\pi)^{-1}r^{(1)}| \\
&= |(I-\gamma P^\pi)^{-1}r^{(1)} + (I-\gamma\widehat{P}^\pi)^{-1}(\gamma\widehat{P}^\pi - \gamma P^\pi)Q^{(1)}| \\
&\lesssim |(I-\gamma P^\pi)^{-1}r^{(1)}| + \frac{\gamma}{\sqrt{N}}|(I-\gamma\widehat{P}^\pi)^{-1}\sqrt{\mathrm{Var}_{\mathrm{P}}(V^{(1)})}| \\
&= |(I-\gamma P^\pi)^{-1}r^{(1)}| + \frac{\gamma}{\sqrt{N}}|(I-\gamma\widehat{P}^\pi)^{-1}r^{(2)}| \\
&\lesssim \cdots \\
&\lesssim \sum_{i=1}^{n}(\frac{\gamma}{\sqrt{N}})^{i-1}|(I-\gamma P^\pi)^{-1}r^{(i)}| + (\frac{\gamma}{\sqrt{N}})^n|(I-\gamma\widehat{P}^\pi)^{-1}r^{(n+1)}|
\end{aligned}$$

In addition, by Lemma B, we have

$$\begin{aligned}
\left|(I-\gamma P^\pi)^{-1}r^{(i)}\right| &= \left|(I-\gamma P^\pi)^{-1}\sqrt{\mathrm{Var}_{\mathrm{P}}(V^{(i-1)})}\right| \\
&\le \frac{4}{\gamma\sqrt{1-\gamma}}|Q^{i-1}| \\
&\le \frac{4}{\gamma\sqrt{1-\gamma}}\left|(I-\gamma\widehat{P}^\pi)^{-1}r^{(i-1)}\right|.
\end{aligned}$$

By induction, we have $\left|(I-\gamma P^\pi)^{-1}r^{(i)}\right| \le \left(\frac{4}{\gamma\sqrt{1-\gamma}}\right)^i\left|(I-\gamma\widehat{P}^\pi)^{-1}r^{(0)}\right| \le \left(\frac{4}{\gamma\sqrt{1-\gamma}}\right)^i\frac{1}{1-\gamma}$. If we set $N \gtrsim \frac{64}{1-\gamma}$ and $n \gtrsim \log(\frac{1}{1-\gamma})$, we have

$$\left|(I-\gamma\widehat{P}^\pi)^{-1}\sqrt{\mathrm{Var}_{\mathrm{P}}(V^\pi)}\right| \le \frac{4}{\gamma\sqrt{(1-\gamma)^3}}\sum_{i=1}^{n}\left(\frac{4}{\sqrt{N(1-\gamma)}}\right)^{i-1} + \frac{1}{(1-\gamma)^2}\left(\frac{4}{\sqrt{N(1-\gamma)}}\right)^n \le \frac{16}{\sqrt{(1-\gamma)^3}}$$

For a more detailed proof on lower order term, one can refer to [Li et al., 2020]. □

B PROBLEM-DEPENDENT UPPER BOUND

Proof of Lemma 4. First, we show that choosing $u = u^*$ in absorbing TBSG can recover $\widehat{Q}^*$.

$$\begin{aligned}
\widetilde{Q}^{\widehat{\pi}^*}_{u^*} &= (I-\gamma\widetilde{P}^{\widehat{\pi}^*})^{-1}(r + (u^* - r(s,a))\mathbf{1}_{s,a}) \\
&= (I-\gamma\widetilde{P}^{\widehat{\pi}^*})^{-1}((I-\gamma\widehat{P}^{\widehat{\pi}^*})\widehat{Q}^* + \gamma(\widehat{P}\widehat{V}^* - \widetilde{P}\widehat{V}^*)) \\
&= \widehat{Q}^*.
\end{aligned}$$

By the property of Nash equilibrium strategy, we have $\widehat{Q}^* = \widetilde{Q}^{\widehat{\pi}^*}_{u^*} = \widetilde{Q}^*_{u^*}$. Specifically, $\widehat{\pi}^*$ select the optimal value in $\widehat{Q}^*$, which means it select the optimal value in $\widetilde{Q}^{\widehat{\pi}^*}_{u^*}$, and this is equivalent to $\widehat{\pi}^* = \widetilde{\pi}^*_{u^*}$.

Second, we show that the optimal Q-value in absorbing TBSG is $\frac{1}{1-\gamma}$-lipschitz to u. This is proved by bounding the distance between an upper bound and an lower bound.

$$\begin{aligned}
\left|\widetilde{Q}^{\widetilde{\mu}^*_u,\widetilde{\nu}^*_{u'}}_u - \widetilde{Q}^{\widetilde{\mu}^*_u,\widetilde{\nu}^*_{u'}}_{u'}\right| &= \left|(I-\widetilde{P}^{\widetilde{\mu}^*_u,\widetilde{\nu}^*_{u'}})^{-1}(r_u - r_{u'})\right| \\
&= \left|(u-u')(I-\widetilde{P}^{\widetilde{\mu}^*_u,\widetilde{\nu}^*_{u'}})^{-1}\mathbf{1}_{s,a}\right| \\
&\le \frac{|u-u'|}{1-\gamma}.
\end{aligned}$$

Similarly, we have $\left|\widetilde{Q}_u^{\widetilde{\mu}_{u'}^*,\widetilde{\nu}_u^*} - \widetilde{Q}_{u'}^{\widetilde{\mu}_{u'}^*,\widetilde{\nu}_u^*}\right| \leq \frac{|u-u'|}{1-\gamma}$. By the definition of Q^*, we have $\widetilde{Q}_u^{\widetilde{\mu}_{u'}^*,\widetilde{\nu}_u^*} \leq \widetilde{Q}_u^* \leq \widetilde{Q}_u^{\widetilde{\mu}_u^*,\widetilde{\nu}_{u'}^*}$ and $\widetilde{Q}_{u'}^{\widetilde{\mu}_u^*,\widetilde{\nu}_{u'}^*} \leq \widetilde{Q}_{u'}^* \leq \widetilde{Q}_{u'}^{\widetilde{\mu}_{u'}^*,\widetilde{\nu}_u^*}$. Thus we have $|\widetilde{Q}_u^* - \widetilde{Q}_{u'}^*| \leq \frac{|u-u'|}{1-\gamma}$. The proof for counterstrategy is the same.

□

Lemma 13. *For any fixed value vector V, with probability larger than $1-\delta$, we have*

$$|(P(s,a) - \widehat{P}(s,a))V| \leq \sqrt{2\log(4/\delta)}\sqrt{\frac{Var_{s,a}(V)}{N}} + \frac{2\log(4/\delta)}{3(1-\gamma)N},$$

where N is the sample size from distribution $P(\cdot|s,a)$.

Proof. The proof is a direct application of Bernstein's inequality. □

Lemma 14. *For $C = 2\log(\frac{32}{(1-\gamma)^2\epsilon\delta})$, with probability larger than $1-\delta$, we have*

$$\left|(P(s,a) - \widehat{P}(s,a))\widehat{V}^{\mu^*,*}\right| \leq \sqrt{\frac{C\mathrm{Var}_{s,a}(\widehat{V}^{\mu^*,*})}{N}} + \frac{C}{3(1-\gamma)N} + \left(\sqrt{\frac{C}{N}} + 1\right)\frac{\epsilon(1-\gamma)}{4},$$

where N is the sample size from distribution $P(\cdot|s,a)$.

Proof. We define a fixed set U to be evenly spaced points in $[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}]$ such that $|U| = \frac{8}{(1-\gamma)^2\epsilon}$. Combining Lemma with union bound, with probability larger than $1-\delta$, we have for all $u \in U$,

$$\left|(P(s,a) - \widehat{P}(s,a))\widetilde{V}_u^{\mu^*,*}\right| \leq \sqrt{2\log(4|U|/\delta)}\sqrt{\frac{Var_{s,a}(\widetilde{V}_u^{\mu^*,*})}{N}} + \frac{2\log(4|U|/\delta)}{3(1-\gamma)N}.$$

For each $u \in U$, we have

$$\begin{aligned}
\left|(P(s,a) - \widehat{P}(s,a))\widehat{V}^{\mu^*,*}\right| &= \left|(P(s,a) - \widehat{P}(s,a))\widetilde{V}_{u^{\mu^*}}^{\mu^*,*}\right| \\
&\leq \left|(P(s,a) - \widehat{P}(s,a))\widetilde{V}_u^{\mu^*,*}\right| + \frac{|u-u^{\mu^*}|}{1-\gamma} \\
&\leq \sqrt{2\log(4|U|/\delta)}\sqrt{\frac{Var_{s,a}(\widetilde{V}_u^{\mu^*,*})}{N}} + \frac{2\log(4|U|/\delta)}{3(1-\gamma)N} + \frac{|u-u^{\mu^*}|}{1-\gamma} \\
&\leq \sqrt{2\log(4|U|/\delta)}\sqrt{\frac{Var_{s,a}(\widetilde{V}_{u^{\mu^*}}^{\mu^*,*})}{N}} + \frac{2\log(4|U|/\delta)}{3(1-\gamma)N} + \left(1 + \sqrt{\frac{2\log(4|U|/\delta)}{N}}\right)\frac{|u-u^{\mu^*}|}{1-\gamma}
\end{aligned}$$

As there exist a $u \in U$ such that $|u - u^{\mu^*}| \leq \frac{\epsilon(1-\gamma)^2}{4}$, we can prove the argument. □

Proof of Lemma 5. By the property of counterstrategy, we have

$$\begin{aligned}
Q^* - \widehat{Q}^* &= Q^* - Q^{\widehat{c}(\nu^*),\nu^*} + Q^{\widehat{c}(\nu^*),\nu^*} - \widehat{Q}^{*,\nu^*} + \widehat{Q}^{*,\nu^*} - \widehat{Q}^* \\
&\geq Q^{\widehat{c}(\nu^*),\nu^*} - \widehat{Q}^{*,\nu^*}.
\end{aligned}$$

Similarly, we have $Q^* - \widehat{Q}^* \leq Q^{\mu^*,\widehat{c}(\mu^*)} - \widehat{Q}^{\mu^*,*}$. Together we can prove Lemma 5. □

Proof of Lemma 6. By the definition of suboptimality gap, for all $s \in \mathcal{S}_{\max}, a \in \mathcal{A}$ we have

$$\begin{aligned}
\widehat{Q}^*(s,\mu^*(s)) - \widehat{Q}^*(s,a) &= \widehat{Q}^*(s,\mu^*(s)) - Q^*(s,\mu^*(s)) + Q^*(s,\mu^*(s)) - Q^*(s,a) + Q^*(s,a) - \widehat{Q}^*(s,a) \\
&\geq -\frac{\Delta}{2} + \Delta - \frac{\Delta}{2} \\
&\geq 0.
\end{aligned}$$

Similar, for all $s \in \mathcal{S}_{\min}, a \in \mathcal{A}$, we have

$$\widehat{Q}^*(s, \nu^*(s)) - \widehat{Q}^*(s, a) < 0.$$

Note that the empirical Nash equilibrium strategy $\widehat{\pi}^*$ is the only policy that can satisfy the above two conditions, which means $\pi^* = \widehat{\pi}^*$. □

Proof of Theorem 1. We set $N = \frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2} \log(\frac{32}{(1-\gamma)^2\epsilon\delta})$. With probability larger than $1-\delta$, We have

$$\begin{aligned}
&\left|Q^{\mu^*,\widehat{c}(\mu^*)} - \widehat{Q}^{\mu^*,*}\right| \\
=&\left|\gamma(I - \gamma P^{\mu^*,\widehat{c}(\mu^*)})^{-1}(P - \widehat{P})\widehat{V}^{\mu^*,\widehat{c}(\mu^*)}\right| \\
\leq&\left|\gamma(I - \gamma P^{\mu^*,\widehat{c}(\mu^*)})^{-1}\left\{\sqrt{\frac{C\mathrm{Var}_{\mathrm{P}}(\widehat{V}^{\mu^*,*})}{N/|\mathcal{S}||\mathcal{A}|}} + \left[\frac{C}{3(1-\gamma)N/|\mathcal{S}||\mathcal{A}|} + \left(\sqrt{\frac{C}{N/|\mathcal{S}||\mathcal{A}|}} + 1\right)\frac{\epsilon(1-\gamma)}{4}\right]\mathbf{1}\right\}\right| \\
\leq&\sqrt{\frac{2C}{(1-\gamma)^3 N/|\mathcal{S}||\mathcal{A}|}} + \sqrt{\frac{C}{N/|\mathcal{S}||\mathcal{A}|}}\frac{\left|Q^{\mu^*,\widehat{c}(\mu^*)} - \widehat{Q}^{\mu^*,*}\right|}{1-\gamma} + \frac{C}{3(1-\gamma)^2 N/|\mathcal{S}||\mathcal{A}|} + \left(\sqrt{\frac{C}{N/|\mathcal{S}||\mathcal{A}|}} + 1\right)\frac{\epsilon}{4}
\end{aligned}$$

where the first equality is due to Lemma 1, the first inequality is due to Lemma 14, the second inequality is due to Lemma 2.

Thus for $\epsilon \leq \frac{1}{\sqrt{1-\gamma}}$ and $N \geq \frac{32|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2} \log(\frac{32}{(1-\gamma)^2\epsilon\delta})$, we have

$$\left|Q^{\mu^*,\widehat{c}(\mu^*)} - \widehat{Q}^{\mu^*,*}\right| \leq \epsilon.$$

Similarly, we have $|Q^{\widehat{c}(\nu^*),\nu^*} - \widehat{Q}^{*,\nu^*}| \leq \epsilon$. Finally, we set $\epsilon = \frac{\Delta}{2}$. Then by Lemma 5 and Lemma 6, we can conclude that with probability $1-\delta$ and $N = \frac{128|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\Delta^2} \log(\frac{128}{(1-\gamma)^2\Delta\delta})$, the empirical Nash equilibrium strategy $\widehat{\pi}^*$ is exactly the true Nash equilibrium strategy π^*.

□

C PROBLEM-INDEPENDENT UPPER BOUND

Proof of Lemma 8. As $Q^*_\tau = \max_\pi Q^\pi_\tau = \max_\pi (I - \gamma P^\pi)^{-1} r_\tau$ is a continuous function of τ, with a tie breaking rule to select optimal action if there are multiple optimal actions, the optimal policy π^*_τ is a piecewise constant function.

For a given policy π, the corresponding Q-function satisfies $Q^\pi_\tau = (I - \gamma P^\pi)^{-1}(r + \tau \mathbf{1}_{s,a})$, which means Q^π_τ is a linear function of τ. As π^*_τ is a piecewise constant function, we have that Q^*_τ is a piecewise linear function.

For the last argument, we factorize $Q^*_\tau(s, a')$ according to the hitting time of state s.

$$\begin{aligned}
Q^*_\tau(s, a') &= r(s, a') + \gamma P(s, a') V^*_\tau \\
&= \gamma p_1 V_\tau(s) + \gamma \sum_{s' \neq s} P(s'|s, a') V^*_\tau(s') + r(s, a') \\
&= \gamma p_1 V_\tau(s) + \gamma^2 p_2 V_\tau(s) + \gamma^2 \sum_{s'', s' \neq s} P(s'|s, a') P(s''|s', \pi^*(s')) V^*_\tau(s'') + r(s, a') + \gamma \sum_{s' \neq s} P(s'|s, a) r(s', \pi^*(s')) \\
&= \cdots \\
&= \sum_{n=1}^{\infty} \gamma^n p_n V^*_\tau(s) + b(\pi^*_\tau),
\end{aligned}$$

where p_n is the probability of first visiting state s in step n under optimal policy and $b(\pi^*_\tau)$ is a function of π^*_τ. For $a = \pi^*_\tau(s)$, we define $k = \sum_{n=1}^{\infty} \gamma^n p_n \leq \gamma$, then $Q^*_\tau(s, a') = k Q^*_\tau(s, a) + b(\pi^*_\tau)$. When $\pi^*_\tau(s) \neq a$, $Q^*_\tau(s, a') = r(s, a') + \gamma P(s, a')(I - \gamma P_{\pi^*_\tau})^{-1} r_{\pi^*_\tau}$ is a function of π^*_τ, which means we have $Q^*_\tau(s, a') = 0 Q^*_\tau(s, a) + b(\pi^*_\tau)$. □

Proof of Lemma 7. We prove that for any state s and actions a, a', with large probability, a gap between $Q^*_{\mathrm{p}}(s,a)$ and $Q^*_{\mathrm{p}}(s,a')$ exist. Then with a union bound, Lemma 7 holds with large probability.

Now we fix all rewards except $r_{\mathrm{p}}(s,a) = r(s,a) + \tau$. We define $\mathcal{I}_w = \{\tau || Q^*_\tau(s,a) - Q^*_\tau(s,a')| \le w\}$. First, we have $Q^*_{\tau_1}(s,a) - Q^*_{\tau_2}(s,a) = \tau_1 - \tau_2 + P(s,a)(V^*_{\tau_1} - V^*_{\tau_2}) \ge \tau_1 - \tau_2, \forall \tau_1 \ge \tau_2$. In addition, Lemma 8 implies that the growth rate of $Q^*_\tau(s,a')$ is at most γ times the rate of $Q_\tau(s,a)$. Thus we can conclude that the length of $\mathcal{I}_w$ is at most $\frac{2w}{1-\gamma}$. We set $w = \frac{\delta\xi(1-\gamma)}{2|\mathcal{S}||\mathcal{A}|^2}$, then gaps between all $Q^*_{\mathrm{p}}(s,a)$ and $Q^*_{\mathrm{p}}(s,a')$ exist, which implies Lemma 7. The proof for counterstrategy is the same. □

Proof of Lemma 9. Select $u \in U$ such that $u - u^* \le \frac{\delta\xi(1-\gamma)^2}{4|\mathcal{S}||\mathcal{A}|^2}$. Thus we have $|\widehat{Q}^*_{\mathrm{p}} - \widetilde{Q}^*_{\mathrm{p},u}| \le \frac{\delta\xi(1-\gamma)}{4|\mathcal{S}||\mathcal{A}|^2}$. As with probability $1-\delta$, a gap of $\frac{\delta\xi(1-\gamma)}{2|\mathcal{S}||\mathcal{A}|^2}$ exist in $\widehat{Q}^*_{\mathrm{p}}$, by Lemma 6, we have $\widehat{\pi}^*_{\mathrm{p}} = \widetilde{\pi}^*_{\mathrm{p},u}$. The proof for $\widehat{c}_{\mathrm{p}}(\mu^*)$ is the same. □

Proof of Lemma 10. First, we consider the unperturbed case.

$$\begin{aligned}
0 &\le Q^* - Q^{\widehat{\mu}^*,*} \\
&= Q^* - Q^{\mu^*,\widehat{c}(\mu^*)} + Q^{\mu^*,\widehat{c}(\mu^*)} - \widehat{Q}^{\mu^*,*} + \widehat{Q}^{\mu^*,*} - \widehat{Q}^* + \widehat{Q}^* - \widehat{Q}^{\widehat{\mu}^*,c(\widehat{\mu}^*)} + \widehat{Q}^{\widehat{\mu}^*,c(\widehat{\mu}^*)} - Q^{\widehat{\mu}^*,*} \\
&\le Q^{\mu^*,\widehat{c}(\mu^*)} - \widehat{Q}^{\mu^*,*} + \widehat{Q}^{\widehat{\mu}^*,c(\widehat{\mu}^*)} - Q^{\widehat{\mu}^*,*}.
\end{aligned}$$

Then we show that the perturbation only induce an error of $\frac{\xi}{1-\gamma}$.

$$\begin{aligned}
\left|Q^{\mu^*,\nu^*_{\mathrm{p}}} - Q^{\mu^*,\nu^*_{\mathrm{p}}}_{\mathrm{p}}\right| &= \left|(I - \gamma P^{\mu^*,\nu^*_{\mathrm{p}}})^{-1}(r - r_{\mathrm{p}})\right| \\
&\le \frac{|r - r_{\mathrm{p}}|}{1-\gamma} \\
&\le \frac{\xi}{1-\gamma}
\end{aligned}$$

Similarly, we have $\left|Q^{\mu^*_{\mathrm{p}},\nu^*} - Q^{\mu^*_{\mathrm{p}},\nu^*}_{\mathrm{p}}\right| \le \frac{\xi}{1-\gamma}$. As $Q^{\mu^*_{\mathrm{p}},\nu^*} \le Q^* \le Q^{\mu^*,\nu^*_{\mathrm{p}}}$ and $Q^{\mu^*_{\mathrm{p}},\nu^*}_{\mathrm{p}} \ge Q^*_{\mathrm{p}} \ge Q^{\mu^*,\nu^*_{\mathrm{p}}}_{\mathrm{p}}$, we have $|Q^*_{\mathrm{p}} - Q^*| \le \frac{\xi}{1-\gamma}$. Thus $|Q^*_{\mathrm{p}} - Q^*| \le \frac{\xi}{1-\gamma}$. Similarly, we have $|Q^{\widehat{\mu}^*,*}_{\mathrm{p}} - Q^{\widehat{\mu}^*,*}| \le \frac{\xi}{1-\gamma}$. Combining all parts together, we have

$$\begin{aligned}
\left|Q^{\widehat{\mu}^*,*} - Q^*\right| &\le \left|Q^{\widehat{\mu}^*,*} - Q^{\widehat{\mu}^*,*}_{\mathrm{p}}\right| + \left|Q^* - Q^*_{\mathrm{p}}\right| + \left|Q^{\widehat{\mu}^*,*}_{\mathrm{p}} - Q^*_{\mathrm{p}}\right| \\
&\le \frac{2\xi}{1-\gamma} + \left|Q^{\mu^*_{\mathrm{p}},\widehat{c}_{\mathrm{p}}(\mu^*_{\mathrm{p}})}_{\mathrm{p}} - \widehat{Q}^{\mu^*_{\mathrm{p}},*}_{\mathrm{p}}\right| + \left|\widehat{Q}^{\widehat{\mu}^*_{\mathrm{p}},c_{\mathrm{p}}(\widehat{\mu}^*_{\mathrm{p}})}_{\mathrm{p}} - Q^{\widehat{\mu}^*_{\mathrm{p}},*}_{\mathrm{p}}\right|
\end{aligned}$$

Combining these inequalities and we can get the proof.

□

Proof of Theorem 2. By Lemma 1, we have

$$Q^{\mu^*_{\mathrm{p}},\widehat{c}_{\mathrm{p}}(\mu^*_{\mathrm{p}})}_{\mathrm{p}} - \widehat{Q}^{\mu^*_{\mathrm{p}},*}_{\mathrm{p}} = \gamma(I - \gamma\widehat{P}^{\mu^*_{\mathrm{p}},\widehat{c}_{\mathrm{p}}(\mu^*_{\mathrm{p}})})^{-1}(P - \widehat{P})V^{\mu^*_{\mathrm{p}},\widehat{c}_{\mathrm{p}}(\mu^*_{\mathrm{p}})}_{\mathrm{p}}$$

By uniform bound, with probability $1-\delta$ for all $u \in U$ defined in Lemma 9, we have

$$\left|(P(s,a) - \widehat{P}(s,a))V^{\mu^*_{\mathrm{p}},\widetilde{c}_{\mathrm{p},u}(\mu^*_{\mathrm{p}})}_{\mathrm{p}}\right| \le \sqrt{2\log(4|U|/\delta)}\sqrt{\frac{Var_{s,a}(V^{\mu^*_{\mathrm{p}},\widetilde{c}_{\mathrm{p},u}(\mu^*_{\mathrm{p}})}_{\mathrm{p}})}{N/|\mathcal{S}||\mathcal{A}|}} + \frac{2\log(4|U|/\delta)}{3(1-\gamma)N/|\mathcal{S}||\mathcal{A}|}.$$

Condition on the event defined in Lemma 9, with probability $1-2\delta$, we have

$$\left|(P(s,a)-\widehat{P}(s,a))V_{\mathrm{p}}^{\mu_{\mathrm{p}}^*,\widehat{c}_{\mathrm{p}}(\mu_{\mathrm{p}}^*)}\right| \leq \sqrt{2\log(4|U|/\delta)}\sqrt{\frac{Var_{s,a}(V_{\mathrm{p}}^{\mu_{\mathrm{p}}^*,\widehat{c}_{\mathrm{p}}(\mu_{\mathrm{p}}^*)})}{N/|\mathcal{S}||\mathcal{A}|}}+\frac{2\log(4|U|/\delta)}{3(1-\gamma)N/|\mathcal{S}||\mathcal{A}|}.$$

With Lemma 3, we have

$$\left|Q_{\mathrm{p}}^{\mu_{\mathrm{p}}^*,\widehat{c}_{\mathrm{p}}(\mu_{\mathrm{p}}^*)}-\widehat{Q}_{\mathrm{p}}^{\mu_{\mathrm{p}}^*,*}\right| \leq \frac{512\log(4|U|/\delta)}{(1-\gamma)^3N/|\mathcal{S}||\mathcal{A}|}+\frac{2\log(4|U|/\delta)}{3(1-\gamma)^2N/|\mathcal{S}||\mathcal{A}|}$$

If we set $N=\frac{8096|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\log(\frac{256|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^2\xi\delta})$, then with probability larger than $1-\delta/2$, we have

$$\left|Q_{\mathrm{p}}^{\mu_{\mathrm{p}}^*,\widehat{c}_{\mathrm{p}}(\mu_{\mathrm{p}}^*)}-\widehat{Q}_{\mathrm{p}}^{\mu_{\mathrm{p}}^*,*}\right| \leq \frac{\epsilon}{4}.$$

Similarly, with probability larger than $1-\delta/2$, we have

$$\left|\widehat{Q}_{\mathrm{p}}^{\widehat{\mu}_{\mathrm{p}}^*,c_{\mathrm{p}}(\widehat{\mu}_{\mathrm{p}}^*)}-Q_{\mathrm{p}}^{\widehat{\mu}_{\mathrm{p}}^*,*}\right| \leq \frac{\epsilon}{4}.$$

Using Lemma 10, with probability larger than $1-\delta$, we have

$$\left|Q^{\widehat{\mu}^*,*}-Q^*\right| \leq \frac{2\xi}{1-\gamma}+\frac{\epsilon}{2}$$

Setting $\xi=\frac{\epsilon(1-\gamma)}{2}$, we have $|Q^{\widehat{\mu}^*,*}-Q^*|\leq\epsilon$. Similarly we can prove $|Q^{*,\widehat{\nu}^*}-Q^*|\leq\epsilon$. Finally we have that if $N\geq\frac{8096|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\log(\frac{256|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^3\epsilon\delta})$, $\widehat{\pi}_{\mathrm{p}}^*=(\widehat{\mu}_{\mathrm{p}}^*,\widehat{\nu}_{\mathrm{p}}^*)$ is an ϵ-Nash equilibrium strategy.

□

Proof of Theorem 3. Set $\epsilon=\frac{\Delta}{2}$ and by Theorem 2, we have that if $N=\frac{C}{(1-\gamma)^3\Delta^2}|\mathcal{S}||\mathcal{A}|\log(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\epsilon})$, then $\widehat{\pi}^*$ is an $\frac{\Delta}{2}$-optimal strategy, which means

$$|Q^*-Q^{\widehat{\pi}^*}|\leq\frac{\Delta}{2}.$$

By the definition of suboptimality gap, we have

$$\forall s\in\mathcal{S}_{\max},a\neq\mu^*(s):Q^*(s,\mu^*(s))-Q^*(s,a)\geq\Delta,$$
$$\forall s\in\mathcal{S}_{\min},a\neq\nu^*(s):Q^*(s,\nu^*(s))-Q^*(s,a)\leq-\Delta.$$

As $|Q^*-Q^{\widehat{\pi}^*}|\leq\frac{\Delta}{2}$, we have

$$\forall s\in\mathcal{S}_{\max},a\neq\mu^*(s):Q^{\widehat{\pi}^*}(s,\mu^*(s))-Q^{\widehat{\pi}^*}(s,a)\geq 0,$$
$$\forall s\in\mathcal{S}_{\min},a\neq\nu^*(s):Q^{\widehat{\pi}^*}(s,\nu^*(s))-Q^{\widehat{\pi}^*}(s,a)\leq 0,$$

which means $\widehat{\pi}^*=\pi^*$.

□