Feature-Based Q-Learning for Two-Player Stochastic Games

Zeyu Jia
Peking University
jiazy@pku.edu.cn

Lin F. Yang
UCLA
linyang@ee.ucla.edu

Mengdi Wang
Princeton University
mengdiw@princeton.edu

Abstract

Consider a two-player zero-sum stochastic game where the transition function can be embedded in a given feature space. We propose a two-player Q-learning algorithm for approximating the Nash equilibrium strategy via sampling. The algorithm is shown to find an ϵ-optimal strategy using sample size linear to the number of features. To further improve its sample efficiency, we develop an accelerated algorithm by adopting techniques such as variance reduction, monotonicity preservation and two-sided strategy approximation. We prove that the algorithm is guaranteed to find an ϵ-optimal strategy using no more than $\tilde{\mathcal{O}}(K/(\epsilon^2(1-\gamma)^4))$ samples with high probability, where K is the number of features and γ is a discount factor. The sample, time and space complexities of the algorithm are independent of original dimensions of the game.

1 Introduction

Two-player turn based stochastic game (2-TBSG) is a generalization of Markov decision process (MDP), both of which are widely used models in machine learning and operations research. While MDP involves one agent with its simple objective to maximize the total reward, 2-TBSG is a zero-sum game involving two players with opposite objectives: one player seeks to maximize the total reward and the other player seeks to minimize the total reward. In a 2-TBSG, the set of all states is divided into two subsets that are controlled by the two players, respectively. We focus on the discounted stationary 2-TBSG, where the probability transition model is invariant across time and the total reward is the infinite sum of all discounted rewards. Our goal is to approximate the Nash equilibrium of the 2-TBSG, whose existence is proved in Shapley (1953).

There are two practical obstacles standing in solving 2-TBSG:

- We usually do not know the transition probability model explicitly;
- The number of possible states and actions are very large or even infinite.

In this paper we have access to a sampling oracle that can generate sample transitions from any state and action pair. We also suppose that a finite number of state-action features are available, such that the unknown probability transition model can be embedded using the feature space. These features allow us to solve 2-TBSG of arbitrary dimensions using parametric algorithms.

A question is raised naturally, that is, how many samples are needed in order to find an approximate Nash equilibrium? For solving the one-player MDP to ϵ-optimality using K features, Yang and Wang (2019) proves an information-theoretic lower bound of sample complexity $\Omega(K/((1-\gamma)^3\epsilon^2))$. Since MDP is a special case of 2-TBSG, the same lower bound applies to 2-TBSG. Yet there has not been any provably efficient algorithm for solving 2-TBSG using features.

To answer this question, we propose two sampling-based algorithms and provide sample complexity analysis. Motivated by the value iteration and Q-learning like algorithms given by Hansen et al. (2013); Yang and Wang (2019), we propose a simple two-player Q-learning algorithm for solving

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

2-TBSG using given features. When the true transition model can be fully embedded in the feature space without losing any information, our algorithm finds an ϵ-optimal strategy using no more than $\tilde{O}(K/((1-\gamma)^7\epsilon^2))$ sample transitions, where K is the number of state-action features. The algorithm, together with its analysis, can be viewed in the appendix (Section B).

To further improve the sample complexity, we use a variance reduction technique, together with a specifically designed monotonicity preservation technique which were previously used in Yang and Wang (2019), to develop an algorithm that is even more sample-efficient. This algorithm uses a two-sided approximation scheme to find the equilibrium value from both above and below. It computes the final ϵ-optimal strategy by sticking two approximate strategies together. This algorithm is proved to find an ϵ-optimal strategy with high probability using $\tilde{O}(K/((1-\gamma)^4\epsilon^2))$ samples, which improves significantly from our first result. Our results are the first and sharpest sample complexity bounds for solving two-player stochastic game using features, to our best knowledges. Our algorithms are the first ones of their kind with provable sample efficiency. It is also worth noting that the algorithms are space and time efficient, whose complexities depend polynomially on K and $\frac{1}{1-\gamma}$, and are free from the game's dimensions. Moreover, we also provide model misspecification error bound for the case where the features cannot fully embed the transition model.

Section 2 presents the problem formulation and basics. The accelerated two-player Q-learning algorithm and its analysis are presented in Section 3 and Section 4. The related literatures has been deferred to the appendix (Section A).

2 Preliminaries

Basics of 2-TBSG A discounted 2-TBSG (2-TBSG for short) consists of a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2, \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$ and $\mathcal{S}_1, \mathcal{S}_2, \mathcal{A}_1, \mathcal{A}_2$ are state sets and action sets for Player 1 and Player 2, respectively. $P \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|}$ is a transition probability matrix, where $P(s'|s,a)$ denotes the probability of transitioning to state s' from state s if action a is used. $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ is the reward vector, where $r(s,a) \in [0,1]$ denotes the immediate reward received using action a at state s.

For a given state $s \in \mathcal{S}$, we use $\mathcal{A}_s$ to denote the available action set for state s. A value function is a mapping from $\mathcal{S}$ to $\mathbb{R}$. A deterministic strategy (strategy for short) $\pi = (\pi_1, \pi_2)$ is defined such that π_1, π_2 are mappings from $\mathcal{S}_1$ to $\mathcal{A}_1$ and from $\mathcal{S}_2$ to $\mathcal{A}_2$, respectively. Given a strategy π, the *value function* of π is defined to be the expectation of total discounted reward starting from s, i.e.,

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r(s_i, \pi(s_i)) \middle| s_0 = s\right], \quad \forall s \in \mathcal{S}, \tag{1}$$

where $\gamma \in [0,1)$ is the discounted factor, and the expectation is over all trajectories starting from s.

Two players in a 2-TBSG has opposite objectives. While the first player seeks to maximize the value function (1), the second player seeks to minimize it. In the following we present the definition of the equilibrium strategy.

Definition 1. A strategy $\pi^* = (\pi_1^*, \pi_2^*)$ is called a Nash equilibrium strategy (equilibrium strategy for short), if $V^{\pi_1,\pi_2^*} \le V^{\pi^*} \le V^{\pi_1^*,\pi_2}$ for any player 1's strategy π_1 and player 2's strategy π_2.

The existence of the Nash equilibrium strategy is proved in Shapley (1953). And all equilibrium strategies share the same value function, which we denote by $v^* \in \mathbb{R}^{|\mathcal{S}|}$.

Notice that v^* is the equilibrium value if and only if it satisfies the following Bellman equation Hansen et al. (2013):

$$v^* = \mathcal{T}v^*, \tag{2}$$

where $\mathcal{T}$ is an operator mapping a value function V into another:

$$\mathcal{T}V(s) = \begin{cases} \max_{a\in\mathcal{A}_s}[r(s,a) + \gamma P(\cdot|s,a)^T V], & \forall s \in \mathcal{S}_1, \\ \min_{a\in\mathcal{A}_s}[r(s,a) + \gamma P(\cdot|s,a)^T V], & \forall s \in \mathcal{S}_2. \end{cases} \tag{3}$$

We give definitions of ϵ-optimal values and ϵ-optimal strategies.

Definition 2. We call a value function V an ϵ-optimal value, if $\|V - v^*\|_\infty \le \epsilon$.

Definition 3. We call a strategy $\pi = (\pi_1, \pi_2)$ an ϵ-optimal strategy, if for any $s \in \mathcal{S}$,

$$\max_{\overline{\pi}_1} \left[V^{\overline{\pi}_1, \pi_2}(s) - v^*(s) \right] \leq \epsilon, \quad \min_{\overline{\pi}_2} \left[V^{\pi_1, \overline{\pi}_2}(s) - v^*(s) \right] \geq -\epsilon.$$

Since $\min_{\overline{\pi}_2} V^{\pi_1, \overline{\pi}_2} \leq v^* \leq \max_{\overline{\pi}_1} V^{\overline{\pi}_1, \pi_2}$, the above definition is equivalent to $\| \min_{\overline{\pi}_2} V^{\pi_1, \overline{\pi}_2} - v^* \|_\infty \leq \epsilon$ and $\| \max_{\overline{\pi}_1} V^{\overline{\pi}_1, \pi_2} - v^* \|_\infty \leq \epsilon$.

Features and Probability Transition Model Suppose we have K feature functions $\phi = \{\phi_1, \cdots, \phi_K\}$ mapping from $\mathcal{S} \times \mathcal{A}$ into $\mathbb{R}$. For every state-action pair (s, a), these features give a feature vector

$$\phi(s, a) = [\phi_1(s, a), \cdots, \phi_K(s, a)]^T \in \mathbb{R}^K.$$

Throughout this paper, we focus on 2-TBSG where the probability transition model can be nearly embedded using the features ϕ without losing any information.

Definition 4. We say that the transition model P can be embedded into the feature space ϕ, if there exists functions $\psi_1, \ldots, \psi_K : \mathcal{S} \mapsto \mathbb{R}$ such that

$$P(s'|s, a) = \sum_{k \in [K]} \phi_k(s, a) \psi_k(s'), \quad \forall s' \in \mathcal{S},\ (s, a) \in \mathcal{S} \times \mathcal{A}.$$

The preceding model is closely related to linear model for Q functions. When P can be fully embedded using ϕ, the Q-functions belong to $span\{r, \phi\}$ so we can parameterize them using K-dimensional vectors. Note that the feature representation is only concerned with the probability transition but has nothing to do with the reward function. It is pointed out by Yang and Wang (2019) that having a transition model which can be embedded into ϕ is equivalent to using linear Q-function model with zero Bellman error. In our subsequent analysis, we also provide approximation guarantee when P cannot be fully embedded using ϕ.

It is worth noting that Definition 4 has a kernel interpretation. It is equivalent to that the left singular functions of P belong to the Hilbert space with the kernel function $K((s, a), (s', a')) = \phi(s, a)^T \phi(s', a')$. Our model and method can be viewed as approximating and solving the 2-TBSG in a given kernel space.

Notations For two value functions V_1, V_2, we use $V_1 \leq V_2$ to denote $V_1(s) \leq V_2(s), \forall s \in \mathcal{S}$. We use $\prod_{[a,b]} V(s)$ to denote the projection of $V(s)$ into the interval $[a, b]$. The total variance (TV) distance between two distributions P_1, P_2 on the state space $\mathcal{S}$ is defined as $\|P_1 - P_2\|_{TV} = \sum_{s \in \mathcal{S}} |P_1(s) - P_2(s)|$. And we use $\tilde{\mathcal{O}}(\cdot)$ to hide log factors of $K, L, \epsilon, 1 - \gamma$ and δ.

3 Variance-Reduced Q-Learning for Two-Player Stochastic Games

In this section, we show how to accelerate the two-player Q-learning algorithm and achieve near-optimal sample efficiency. A main technique is to leverage monotonicity of the Bellman operator to guarantee that solutions improve monotonically in the algorithm, which was used in Yang and Wang (2019). The entire algorithm can be viewed in appendix (Section C).

3.1 Nonnegative Features

To preserve monotonicity in the algorithm, we assume without loss of generality the features are nonnegative:

$$\phi_k(s, a) \geq 0, \quad \forall k \in [K], \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

This condition can be easily satisfied. If the raw features ϕ does not satisfy nonnegativity, we can construct new features ϕ' to make it satisfied. For any state-action pair (s, a) we append another 1D feature $\phi'_{K+1}(s, a)$ such that $\phi'_k(s, a) = \phi_k(s, a) + \phi'_{K+1}(s, a) \geq 0$ for $k \in [K]$, and there is a subset $\mathcal{K}'$ of $\mathcal{S} \times \mathcal{A}$ such that $|\mathcal{K}'| = K + 1$ and $\Phi'_{\mathcal{K}'}$ is nonsingular. Then ϕ' satisfies nonnegativity condition and Assumption 1 for some L by normalization. More details are deferred to appendix.

3.2 Parametrization

We use a “max-linear" parameterization to guarantee that value functions improve monotonically in the algorithm. Instead of using a single vector w, we now use a finite collection of K-dimensional vectors $\theta = \{w^{(h)}\}_{h=1}^{Z}$, where Z is an integer of order $1/(1-\gamma)$. We use the following parameterization for the Q-functions, the value functions and strategies[1]:

$$\begin{aligned}
&Q_{w^{(h)}}(s,a) = r(s,a) + \gamma\phi(s,a)^T w^{(h)},\\
&V_\theta(s) = \begin{cases} \max_{h\in[Z]} \max_{a\in\mathcal{A}_s} Q_{w^{(h)}}(s,a), & \forall s \in \mathcal{S}_1,\\ \max_{h\in[Z]} \min_{a\in\mathcal{A}_s} Q_{w^{(h)}}(s,a), & \forall s \in \mathcal{S}_2,\end{cases}\\
&\pi_\theta(s) = \begin{cases} \max_{h\in[Z]} \left[\arg\max_{a\in\mathcal{A}_s} \left(Q_{w^{(h)}}(s,a)\right)\right], & \forall s \in \mathcal{S}_1,\\ \max_{h\in[Z]} \left[\arg\min_{a\in\mathcal{A}_s} \left(Q_{w^{(h)}}(s,a)\right)\right], & \forall s \in \mathcal{S}_2.\end{cases}
\end{aligned} \tag{4}$$

For a given θ and s, computing the corresponding Q-value and action requires solving a one-step optimization problem. We assume that there is an oracle that solves the problem with time complexity M.

Remark 1. When the action space is continuous, this M may become a constant which is independent to the state set and the action set.

3.3 Preserving Monotonicity

A drawback of value iteration-like method is: an ϵ-optimal value function cannot be used greedily to obtain an ϵ-optimal strategy. In order for Algorithm 1 to find an π such that V^π is an ϵ-optimal value, it needs to find an $(1-\gamma)\epsilon$-optimal value function first, which is very inefficient. However, if a strategy π and a value function V satisfy following inequality

$$V \le \mathcal{T}_\pi V, \tag{5}$$

then there is a strong connection between V^π and V as follows (due to monotonicity of the Bellman operator $\mathcal{T}$):

$$V \le \mathcal{T}_\pi V \le \mathcal{T}_\pi^2 V \le \cdots \le \mathcal{T}_\pi^\infty V = V^\pi.$$

This relation will be used to show that if V is close to optimal, the policy π is also close to optimal.

The accelerated algorithm is given partly in Algorithm 2, which uses two tricks to preserve monotonicity:

- We use parametrization (4) for Q_w, V_θ and π_θ. This parametrization ensures that in our algorithm, the values and strategies keeps improving throughout iterations.
- In each iteration, we shift downwards the new parameter $w^{(i,j)}$ to $\overline{w}^{(i,j)}$ by using a confidence bound, such that

$$\phi(s,a)^T \overline{w}^{(i,j)} \le P(\cdot|s,a)^T V^{(i,j-1)} \le P(\cdot|s,a)^T V^{(i,j)},$$

  which uses the nonnegativity of features. The shift is used to guarantee (5).

3.4 Approximating the Equilibrium from Two Sides

Making value functions monotonically increasing is not enough to find an ϵ-optimal strategy for two-player stochastic games. There are two sides of the game, and V^π may be either greater or less than the Nash value. Having a lowerbound V for V^π does not lead to an approximate strategy. This is a major difference from one-player MDP.

In order to fix this problem, we approximate the Nash equilibrium from two sides – both from above and below. Given player 1's strategy π_1 and player 2's strategy π_2, we introduce two Bellman operators $\mathcal{T}_{\pi_1,\min}, \mathcal{T}_{\max,\pi_2}$.

$$\begin{aligned}
&\mathcal{T}_{\pi_1,\min} V = \begin{cases} r(s,\pi_1(s)) + \gamma P(\cdot|s,\pi_1(s))^T V, & \text{if } s \in \mathcal{S}_1,\\ \min_{s\in\mathcal{A}_s}\left[r(s,a) + \gamma P(\cdot|s,a)^T V\right], & \text{if } s\in\mathcal{S}_2,\end{cases}\\
&\mathcal{T}_{\max,\pi_2} V = \begin{cases} \max_{s\in\mathcal{A}_s}\left[r(s,a) + \gamma P(\cdot|s,a)^T V\right], & \text{if } s \in \mathcal{S}_1,\\ r(s,\pi_2(s)) + \gamma P(\cdot|s,\pi_2(s))^T V, & \text{if } s\in\mathcal{S}_2.\end{cases}
\end{aligned} \tag{6}$$

[1]Here $\max_{h\in[Z]}[\arg\min_{a\in\mathcal{A}_s} Q_{w^{(h)}}(s,a)]$ is defined to be the solution of a in the max-min problem: $\max_{h\in[Z]}\min_{a\in\mathcal{A}_s} Q_{w^{(h)}}(s,a)$. The definition of $\max_{h\in[Z]}[\arg\max_{a\in\mathcal{A}_s} Q_{w^{(h)}}(s,a)]$ is similar.

Then if there exist value functions V, W such that all of the following

$$V \le \mathcal{T}_{\pi_1,\min} V, \quad \mathcal{T}_{\max,\pi_2} W \le W \tag{7}$$

hold, then by using the monotonicity of $\mathcal{T}_{\pi_1,\min}, \mathcal{T}_{\pi_2,\max}$ we get

$$V \le \min_{\overline{\pi}_2} V^{\pi_1,\overline{\pi}_2} \le v^* \le \max_{\overline{\pi}_1} V^{\overline{\pi}_1,\pi_2} \le W.$$

Hence if we have $\|V - v^*\|_\infty \le \epsilon$ and $\|W - v^*\|_\infty \le \epsilon$, they jointly imply

$$\begin{aligned} \|\min_{\overline{\pi}_2} V^{\pi_1,\overline{\pi}_2} - v^*\|_\infty &\le \epsilon, \\ \|\max_{\overline{\pi}_1} V^{\overline{\pi}_1,\pi_2} - v^*\|_\infty &\le \epsilon, \end{aligned} \tag{8}$$

which indicates that (π_1, π_2) is an ϵ-optimal strategy.

To achieve this goal, we construct a "flipped" instance of 2-TBSG $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P, r', \gamma)$, where the state set and the action set for each player, the transition probability matrix and the discounted factor are identical with those of $\mathcal{M}$. The reward function r' is defined as

$$r'(s,a) = 1 - r(s,a). \tag{9}$$

And the objective of two players are switched, which means in $\mathcal{M}'$ the first player aims to minimize and the second player aims to maximize. $\mathcal{M}, \mathcal{M}'$ share the same optimal strategy (but flipped).

We use V' to denote the value function of $\mathcal{M}'$, and let $W(s) = \frac{1}{1-\gamma} - V'(s)$ for any $s \in \mathcal{S}$, which serves as the value function approximating the equilibrium value v^* from upper side. This W, together with V, forms a two-sided approximation to the equilibrium value function.

We use Algorithm 2 to solve $\mathcal{M}$ and $\mathcal{M}'$ at the same time. Next we construct a strategy π where the first player's strategy is based on parameters from the lower approximation, and the second player's strategy is based on parameters from the upper approximation. This process is described in Algorithm 3, and its output is the following approximate Nash equilibrium strategy:

$$\pi(s) = \begin{cases} \pi_{\theta^{(R',R)}}(s), & \text{if } s \in \mathcal{S}_1, \\ \pi'_{\eta^{(R',R)}}(s), & \text{if } s \in \mathcal{S}_2, \end{cases} \tag{10}$$

where for $\eta = \{z^{(h)}\}_{h=1}^Z$, π' is the strategy defined as

$$\pi'_\eta(s) = \begin{cases} \max_{h\in[Z]} \left[\arg\min_{a\in\mathcal{A}_s} \left(r'(s,a) + \gamma\phi(s,a)^T z^{(h)}\right)\right], & \forall s \in \mathcal{S}_1, \\ \max_{h\in[Z]} \left[\arg\max_{a\in\mathcal{A}_s} \left(r'(s,a) + \gamma\phi(s,a)^T z^{(h)}\right)\right], & \forall s \in \mathcal{S}_2, \end{cases}$$

3.5 Variance Reduction

We use inner-outer loops for variance reduction in our algorithm. Let the parameters at the (i,j)-th iteration be $\theta^{(i,j)}$. At the beginning of the i-th outer iteration, we aim to approximate $P(\cdot|s,a)^T V_{\theta^{(i,0)}}$ accurately (Step 6, 7). Then in the (i,j)-th inner iteration, we use $P(\cdot|s,a)^T V_{\theta^{(i,0)}}$ as a reference to reduce the variance of estimation. That is, we estimate the difference $P(\cdot|s,a)^T(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}})$ using samples and then use the following equation (Step 11, 12 in Algorithm 2)

$$P(\cdot|s,a)^T V_{\theta^{(i,j-1)}} = P(\cdot|s,a)^T V_{\theta^{(i,0)}} + P(\cdot|s,a)^T(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}})$$

to approximate $P(\cdot|s,a)^T V_{\theta^{(i,j-1)}}$. Since the infinite norm of $(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}})$ is guaranteed to be smaller than the absolute value of $V_{\theta^{(i,0)}}$, the number of samples needed for each inner iteration can be substantially reduced. Hence our algorithm is more sample-efficient.

3.6 Putting Together

Algorithms 2-3 puts together all the techniques that were explained. In the next section, we will prove that they output an ϵ-optimal strategy with high probability. It is easy to see that the time complexity of Algorithm 2 is $\tilde{\mathcal{O}}(K^\omega + MK/((1-\gamma)^4\epsilon^2) + K^2/(1-\gamma))$. The first term K^ω is the time calculating $\Phi_\mathcal{K}^{-1}$. The second term is the time of sampling and calculating the value function in each iteration, and M is the time of calculating $V_\theta(s)$ given parameter θ and state s, which can be viewed as solving an optimization problem over the action space. The last term is due to the calculation of $\Phi_\mathcal{K} M^{(i,j)}$. As for the space complexity, we only need to store $\Phi_\mathcal{K}^{-1}$ and the parameter $\theta^{(i,j)}$ at each iteration, which take $\tilde{\mathcal{O}}(K^2 + K/(1-\gamma))$ space. Hence the total time and space complexities are independent from the numbers of states and actions.

4 Sample Complexity of Algorithms 2-3

In this section, we analyze the sample complexity of our Algorithms 2-3.

Theorem 1. *Let Assumption 1 hold and let features be nonnegative. Suppose that the transition model P of $\mathcal{M}$ can be fully embedded into ϕ space. Then for some $\epsilon > 0, \delta \in (0,1)$, with probability at least $1-2\delta$, the output of Algorithm 3 is an ϵ-optimal strategy. The number of samples used is $N = \mathcal{O}\left(\frac{KL^2}{(1-\gamma)^4\epsilon^2} \cdot poly \log(K\epsilon^{-1}\delta^{-1})\right)$.*

We present a proof sketch here, and the complete proof is deferred to appendix.

Proof Sketch. We prove $\|V_{\theta^{(i,0)}} - v^*\|_\infty \leq 2^{-i}/(1-\gamma)$ by induction. It is easy to know that $\|V_{\theta^{(0,0)}} - v^*\|_\infty \leq 1/(1-\gamma)$. Next we assume $\|V_{\theta^{(i-1,0)}} - v^*\|_\infty \leq 2^{-i+1}/(1-\gamma)$ holds.

The error between $V_{\theta^{(i,0)}}$ and v^* involves two types of error: the estimation error due to sampling and the convergence error of value iteration. Due to the variance reduction technique, estimation error has two parts. The first part is the estimation error of $\Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V_{\theta^{(i-1,0)}}$, which we denote as $\epsilon^{(i,0)}$, and the second part is the error of $\Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}(V_{\theta^{(i-1,j)}} - V_{\theta^{(i-1,0)}})$, which we denote as $\epsilon^{(i,j)}$ for short.

According to the Hoeffding inequality, we have $\epsilon^{(i,0)} = \tilde{\mathcal{O}}(L/(1-\gamma)\cdot\sqrt{1/m})$ and $\epsilon^{(i,j)} = \tilde{\mathcal{O}}(L \cdot \max_s |V_{\theta^{(i-1,j-1)}}(s) - V_{\theta^{(i-1,0)}}(s)| \cdot \sqrt{1/m_1})$ with high probability. By the induction hypothesis, we have $\max_s |V_{\theta^{(i-1,j-1)}}(s) - V_{\theta^{(i-1,0)}}(s)| \leq \frac{1}{2^{i-1}(1-\gamma)}$. If we choose $m = cL^2/((1-\gamma)^4\epsilon^2)$ and $m_1 = c_1 L^2/(1-\gamma)^2$, we will have $\epsilon^{(i,0)} = \mathcal{O}(\epsilon(1-\gamma))$ and $\epsilon^{(i,j)} = \mathcal{O}(2^{-i})$.

The convergence error of value iteration in the inner loop is $\gamma^R/(1-\gamma)$. If we choose $R = c_R \log(\epsilon^{-1}(1-\gamma)^{-1})/(1-\gamma)$, we will have $\gamma^R/(1-\gamma) = \mathcal{O}(\epsilon)$. Bringing these two types of errors together, we have with high probability that

$$\|v^* - V_{\theta^{(i,0)}}\|_\infty \leq \frac{\gamma^R}{1-\gamma} + \sum_{j=1}^{R} \gamma^{R-j}(\epsilon^{(i,0)} + \epsilon^{(i,j)}) \leq \mathcal{O}\left(\epsilon + \frac{\epsilon(1-\gamma) + 2^{-i}}{1-\gamma}\right) = \mathcal{O}\left(\frac{2^{-i}}{1-\gamma}\right),$$

where the last equality is due to $\epsilon \leq 2^{-R'}/(1-\gamma) \leq 2^{-i}/(1-\gamma)$. Choosing $R' = c_{R'} \log(\epsilon^{-1}(1-\gamma)^{-1})$, we have $\|v^* - V_{\theta^{(R',R)}}\|_\infty = \|v^* - V_{\theta^{(R+1,0)}}\|_\infty \leq \epsilon$. Here we have omitted the dependence on any constant factors.

Similarly, we can show $\|v^* - W_{\eta^{(R+1,0)}}\|_\infty \leq \epsilon$ for the "flipped" side. Hence we have $V_{\theta^{(R',R)}} \leq v^* \leq W_{\eta^{(R',R)}}$, therefore the combined strategy $\pi = (\pi_1, \pi_2)$ given by Algorithm 3 is an ϵ-optimal strategy since

$$V_{\theta^{(R',R)}} \leq \min_{\overline{\pi}_2} V^{\pi_1, \overline{\pi}_2} \leq v^* \leq \max_{\overline{\pi}_1} V^{\overline{\pi}_1, \pi_2} \leq W_{\eta^{(R,R')}}, \tag{11}$$

whose proof is based on the monotonicity of two operators $\mathcal{T}_{\pi_1,\min}, \mathcal{T}_{\max,\pi_2}$. The total number of samples used by Algorithm 3 is $R'(R \cdot m_1 + m) = \tilde{\mathcal{O}}(L^2/((1-\gamma)^4\epsilon^2))$. □

According to Theorem 3 and 1, we have the following theorem when the transition model cannot be embedded exactly, whose proof is deferred to appendix.

Theorem 2 (Approximation error due to model misspecification)**.** *Let Assumption 1 holds and let features be nonnegative. If there is an another transition model $\tilde{P}$ which can be fully embedded into ϕ space, and there exists $\xi \in [0,1]$ such that $\|P(\cdot|s,a) - \tilde{P}(\cdot|s,a)\|_{TV} \leq \xi$ for $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ and $P(\cdot|s,a) = \tilde{P}(\cdot|s,a)$ for $(s,a) \in \mathcal{K}$, then with probability at least $1-2\delta$, the output of Algorithm 3 is an $\left(2\xi/(1-\gamma)^2 + 2\epsilon\right)$-optimal strategy, and with probability at least $1-\delta$, the parametrized strategy $(\pi^1_{w^{(R)}}, \pi^2_{w^{(R)}})$ according to the output of Algorithm 1 is $\left(2\xi/(1-\gamma)^2 + 2\epsilon\right)$-optimal.*

5 Conclusion

In this paper, we develop a two-player Q-learning algorithm for solving 2-TBSG in feature space. This algorithm is proved to find an ϵ-optimal strategy with high probability using $\tilde{O}(K/((1-\gamma)^4\epsilon^2))$ samples. It is the first and sharpest sample complexity bound for solving two-player stochastic game using features and linear models, to our best knowledges. The algorithm is sample efficient as well as space and time efficient.

References

Antos, A., Szepesvári, C., and Munos, R. (2008a). Fitted q-iteration in continuous action-space mdps. In *Advances in neural information processing systems*, pages 9–16.

Antos, A., Szepesvári, C., and Munos, R. (2008b). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Azar, M. G., Munos, R., Ghavamzadeh, M., and Kappen, H. (2011a). Reinforcement learning with a near optimal rate of convergence.

Azar, M. G., Munos, R., Ghavamzadeh, M., and Kappen, H. (2011b). Speedy q-learning. In *Advances in neural information processing systems*.

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

Chen, Y., Li, L., and Wang, M. (2018). Scalable bilinear pi learning using state and action features. In *Proceedings of the 35th International Conference on Machine Learning*, pages 834–843, Stockholmsmässan, Stockholm Sweden. PMLR.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105.

Farahmand, A.-m., Szepesvári, C., and Munos, R. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pages 568–576.

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

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

Hu, J. and Wellman, M. P. (1999). Multiagent reinforcement learning in stochastic games. *Submitted for publication*.

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

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713. JMLR. org.

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

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

Lagoudakis, M. G. and Parr, R. (2003). Least-squares policy iteration. *Journal of machine learning research*, 4(Dec):1107–1149.

Lazaric, A., Ghavamzadeh, M., and Munos, R. (2012). Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13(Oct):3041–3074.

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

Littman, M. L. (1996). Algorithms for sequential decision making.

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

Littman, M. L. (2001b). Value-function reinforcement learning in markov games. *Cognitive Systems Research*, 2(1):55–66.

Ludwig, W. (1995). A subexponential randomized algorithm for the simple stochastic game problem. *Information and computation*, 117(1):151–155.

Melo, F. S., Meyn, S. P., and Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671. ACM.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857.

Nedić, A. and Bertsekas, D. P. (2003). Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1-2):79–110.

Neyman, A., Sorin, S., and Sorin, S. (2003). *Stochastic games and applications*, volume 570. Springer Science & Business Media.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., and Littman, M. L. (2008). An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 752–759. ACM.

Pérolat, J., Piot, B., Geist, M., Scherrer, B., and Pietquin, O. (2016). Softened approximate policy iteration for markov games. In *ICML 2016-33rd International Conference on Machine Learning*.

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

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rao, S. S., Chandrasekaran, R., and Nair, K. (1973). Algorithms for discounted stochastic games. *Journal of Optimization Theory and Applications*, 11(6):627–637.

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

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018a). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 5186–5196. Curran Associates, Inc.

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018b). 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. Society for Industrial and Applied Mathematics.

Singh, S. P. and Yee, R. C. (1994). An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233.

Sutton, R. S., Maei, H. R., and Szepesvári, C. (2009). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, pages 1609–1616.

Tagorti, M. and Scherrer, B. (2015). On the rate of convergence and error bounds for LSTD(λ). In *International Conference on Machine Learning*, pages 1521–1529.

Tsitsiklis, J. N. and Van Roy, B. (1997). Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*, pages 1075–1081.

Von Neumann, J. and Morgenstern, O. (2007). *Theory of games and economic behavior (commemorative edition)*. Princeton university press.

Yang, L. F. and Wang, M. (2019). Sample-optimal parametric q-learning with linear transition models. *arXiv preprint arXiv:1902.04779*.

Yang, Z., Xie, Y., and Wang, Z. (2019). A theoretical analysis of deep q-learning. *arXiv preprint arXiv:1901.00137*.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar, T. (2018). Finite-sample analyses for fully decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:1812.02783*.

A Related Works

The 2-TBSG is a special case of games and stochastic games (SG), which are first introduced in Von Neumann and Morgenstern (2007) and Shapley (1953). For a comprehensive introduction on SG, please refer to the books Neyman et al. (2003) and Filar and Vrieze (2012). A number of deterministic algorithms have been developed for solving 2-TBSG when its explicit form is fully given, including Littman (1996); Ludwig (1995); Hansen et al. (2013). For example Rao et al. (1973) proposes the strategy iteration algorithm. A value iteration method is proposed by Hansen et al. (2013), which is one of the motivation of our algorithm.

In the special case of MDP, there exist a large body of works on its sample complexity and sampling-based algorithms. For the tabular setting (finitely many state and actions), sample complexity of MDP with a sampling oracle has been studied in Kearns and Singh (1999); Azar et al. (2013); Sidford et al. (2018b,a); Kakade (2003); Singh and Yee (1994); Azar et al. (2011b). Lower bounds for sample complexity have been studied in Azar et al. (2013); Even-Dar et al. (2006); Azar et al. (2011a), where the first tight lower bound $\Omega(|\mathcal{S}||\mathcal{A}|/(1-\gamma)^3)$ is obtained in Azar et al. (2013). The first sample-optimal algorithm for finding an ϵ-optimal value is proposed in Azar et al. (2013). Sidford et al. (2018a) gives the first algorithm that finds an ϵ-optimal policy using the optimal sample complexity $\tilde{O}(|\mathcal{S}||\mathcal{A}|/(1-\gamma)^3)$ for *all* values of ϵ. For solving MDP using K linearly additive features, Yang and Wang (2019) proved a lower bound of sample complexity that is $\Omega(K/((1-\gamma)^3\epsilon^2))$. It also provided an algorithm that achieves this lower bound up to log factors, however, their analysis of the algorithm relies heavily on an extra "anchor state" assumption. In Chen et al. (2018), a primal-dual method solving MDP with linear and bilinear representation of value functions and transition models is proposed for the undiscounted MDP. In Jiang et al. (2017), the sample complexity of contextual decision process is studied.

As for general stochastic games, the minimax Q-learning algorithm and the friend-and-foe Q-learning algorithm is introduced in Littman (1994) and Littman (2001a), respectively. The Nash Q-learning algorithm is proposed for zero-sum games in Hu and Wellman (2003) and for general-sum games in Littman (2001b); Hu and Wellman (1999). Also in Perolat et al. (2015), the error of approximate Q-learning is estimated. In Zhang et al. (2018), finite-sample analysis of multi-agent reinforcement learning is provided. To our best knowledge, there is no known algorithm that solves 2-TBSG using features with sample complexity analysis.

There are a large number of works analyzing linear model approximation of value and Q functions, for examples Tsitsiklis and Van Roy (1997); Nedić and Bertsekas (2003); Lagoudakis and Parr (2003); Melo et al. (2008); Parr et al. (2008); Sutton et al. (2009); Lazaric et al. (2012); Tagorti and Scherrer (2015). These work mainly focus on approximating the value function or Q function for a fixed policy. The convergence of temporal difference learning with a linear model for a given policy is proved in Tsitsiklis and Van Roy (1997). Melo et al. (2008) and Sutton et al. (2009) study the convergence of Q-learning and off-policy temporal difference learning with linear function parametrization, respectively. In Parr et al. (2008), the relationship of linear transition model and linear parametrized value functions is explained. It is also pointed out by Yang and Wang (2019) that using linear model for Q function is essentially equivalent to assuming that the transition model can be embedded using these features, provided that there is zero Bellman error.

The fitted value iteration for MDPs or 2TBSGs, where the value function is approximated by functions in a general function space, is analyzed in Munos and Szepesvári (2008); Antos et al. (2008a,b); Farahmand et al. (2010); Yang et al. (2019); Pérolat et al. (2016). In these papers, it is shown that the error is related to the Bellman error of the function space, and depends polynomially on $1/\epsilon, 1/(1-\gamma)$ and the dimension of the function space. However, only convergence is analyzed in these paper.

B A Basic Two-Player Q-learning Algorithm

In this section, we develop a basic two-player Q learning algorithm for 2-TBSG. The algorithm is motivated by the two-player value iteration algorithm Hansen et al. (2013). It is also motivated by the parametric Q-learning algorithm for solving MDP given by Yang and Wang (2019).

B.1 Algorithm and Parametrization

The algorithm uses a vector $w \in \mathbb{R}^K$ to parametrize Q-functions, value functions and strategies as follows:

$$
\begin{aligned}
&Q_w(s,a) = r(s,a) + \gamma\phi(s,a)^T w,\\
&V_w(s) = \begin{cases} \max_{a\in\mathcal{A}} Q_w(s,a) & s \in \mathcal{S}_1, \\ \min_{a\in\mathcal{A}} Q_w(s,a) & s \in \mathcal{S}_2, \end{cases} \qquad \pi_w(s) = \begin{cases} \arg\max_{a\in\mathcal{A}} Q_w(s,a) & s \in \mathcal{S}_1, \\ \arg\min_{a\in\mathcal{A}} Q_w(s,a) & s \in \mathcal{S}_2. \end{cases}
\end{aligned} \tag{12}
$$

The algorithm keeps tracks of the parameter vector w only. The value functions and strategies can be obtained from w according to preceding equations when they are needed.

We present Algorithm 1, which is an approximate value iteration. Our algorithm picks a set $\mathcal{K}$ of representative state-action pairs at first. Then at iteration t, it uses sampling to estimate the values $P(\cdot|s,a)^T V_{w^{(t-1)}}$, and carries value iteration using these estimates. The set $\mathcal{K}$ can be chosen nearly arbitrarily, but it is necessary that the set is representative of the feature space. It means that the feature vectors of state-action pairs in this set cannot too be alike but need to be linearly independent.

Assumption 1. There exist K state-action pairs (s,a) forming a set $\mathcal{K}$ satisfying

$$
\|\phi(s,a)\|_1 \le 1, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}_s, \qquad \exists L > 0, \quad \|\Phi_{\mathcal{K}}^{-1}\|_\infty \le L,
$$

where $\Phi_{\mathcal{K}}$ is the $K \times K$ matrix formed by row features of those (s,a) in $\mathcal{K}$.

Algorithm 1 Sampled Value Iteration for 2-TBSG

Input: A 2-TBSG $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$.
Input: $\epsilon, \delta \in (0,1)$.
Initialize: $w^{(0)} \leftarrow 0$.
Initialize: $R \leftarrow \Theta(1/(1-\gamma)\log(1/(1-\gamma)\epsilon))$, $\quad T \leftarrow \Theta\left[\left(L^2 \cdot \log(KR/\delta)/(\epsilon^2(1-\gamma)^6)\right)\right]$.
Pick a set $\mathcal{K}$ of state-action pairs, which satisfies Assumption 1.
for $t = 1 : R$ **do**
  $M^{(t)} \leftarrow \mathbf{0} \in \mathbb{R}^K$
  for $(s,a) \in \mathcal{K}$ **do**
    Sample $P(\cdot|s,a)$ for T times to obtain $s_1^{(t)}, \cdots, s_T^{(t)} \in \mathcal{S}$.
    Let $M^{(t)}(s,a) = \frac{1}{T}\sum_{i=1}^T \Pi_{[0,1/(1-\gamma)]} V_{w^{(t-1)}}(s_i^{(t)})$, where V_w is defined in (12).
  end for
  $w^{(t)} \leftarrow \Phi_{\mathcal{K}}^{-1} M^{(t)}$.
end for
Output: $w^{(R)} \in \mathbb{R}^K$.

B.2 Sample Complexity Analysis

The next theorem establishes the sample complexity of Algorithm 1, which is independent from $|\mathcal{S}|$ and $|\mathcal{A}|$. Its proof is deferred to the appendix.

Theorem 3 (Convergence of Algorithm 1)**.** *Let Assumption 1 holds. Suppose that the transition model P of $\mathcal{M}$ can be fully embedded into the ϕ space. Then for some $\epsilon > 0, \delta \in (0,1)$, with probability at least $1-\delta$, the parametrized strategy $\pi_{w^{(R)}}$ according to the output of Algorithm 1 is ϵ-optimal. The number of samples used is $\tilde{O}\left(\frac{KL^2}{(1-\gamma)^7\epsilon^2} \cdot \text{poly}\log(K\epsilon^{-1}\delta^{-1})\right)$.*

B.3 Proof of Theorem 3

We first present the definition of optimal counterstrategies.

Definition 5. For player 1's strategy π_1, we call π_2 a player 2's optimal counterstrategy against π_1, if for any player 2's strategy $\overline{\pi}_2$, we have $V^{\pi_1,\pi_2} \le V^{\pi_1,\overline{\pi}_2}$. For player 2's strategy π_2, we call π_1 a player 1's optimal counterstrategy against π_2, if for any player 1's strategy $\overline{\pi}_1$, we have $V^{\pi_1,\pi_2} \ge V^{\overline{\pi}_1,\pi_2}$.

It is known in Puterman (2014) that for any player 1's strategy π_1 (player 2's strategy π_2), the optimal counterstrategy against π_1 (π_2) always exists.

Our next lemma indicates that we can use the error of parametrized Q functions to bounded the error of value functions of parametrized strategies.

Lemma 1. *If*

$$\|Q_w(s,a) - Q^*(s,a)\|_\infty \le \zeta, \tag{13}$$

then we have

$$\|v^{\pi_1,\pi_2^*} - v^*\|_\infty \le \frac{2\zeta}{1-\gamma}, \quad \|v^{\pi_1^*,\pi_2} - v^*\|_\infty \le \frac{2\zeta}{1-\gamma}, \tag{14}$$

where $(\pi_1,\pi_2) = \pi_w$*, and* π_1^*, π_2^* *are optimal counterstrategies of* π_2, π_1.

Proof. We only prove the first inequality of (14). The proof of the second inequality is similar.

For any $s \in \mathcal{S}_1$,

$$\begin{aligned}
|v^{\pi_1,\pi_2^*}(s) - v^*(s)| &= |v^{\pi_1,\pi_2^*}(s) - Q^*(s,\pi^*(s))| \\
&\le |v^{\pi_1,\pi_2^*}(s) - Q^*(s,\pi_1(s))| + |Q^*(s,\pi_1(s)) - Q^*(s,\pi^*(s))| \\
&= |\gamma P(\cdot|s,\pi_1(s)) v^{\pi_1,\pi_2^*} - \gamma P(\cdot|s,\pi_1(s)) v^*| + |Q^*(s,\pi_1(s)) - Q^*(s,\pi^*(s))| \\
&\le \gamma |P(\cdot|s,\pi_1(s))(v^{\pi_1,\pi_2^*} - v^*)| + |Q^*(s,\pi_1(s)) - Q^*(s,\pi^*(s))| \\
&\le \gamma \|v^{\pi_1,\pi_2^*} - v^*\|_\infty + |Q^*(s,\pi_1(s)) - Q^*(s,\pi^*(s))|.
\end{aligned}$$

According to the definition of $\pi_w = (\pi_1,\pi_2)$ and Q^*,

$$\begin{aligned}
&Q^*(s,\pi_1(s)) \le Q^*(s,\pi^*(s)), \\
&Q_w(s,\pi_1(s)) \ge Q_w(s,\pi^*(s)).
\end{aligned}$$

Combine this inequality with inequality (13), we get

$$\begin{aligned}
0 &\le Q^*(s,\pi^*(s)) - Q^*(s,\pi_1(s)) \\
&\le |Q^*(s,\pi^*(s)) - Q_w(s,\pi^*(s))| + Q_w(s,\pi^*(s)) - Q_w(s,\pi_1(s)) \\
&\quad + |Q_w(s,\pi_1(s)) - Q^*(s,\pi_1(s))| \\
&\le 2\|Q_w - Q^*\|_\infty \le 2\zeta,
\end{aligned}$$

which indicates that

$$|v^{\pi_1,\pi_2^*}(s) - v^*(s)| \le \gamma \|v^{\pi_1,\pi_2^*} - v^*\|_\infty + 2\zeta, \quad \forall s \in \mathcal{S}_1.$$

Furthermore, for $s \in \mathcal{S}_2$, we have

$$\begin{aligned}
|v^{\pi_1,\pi_2^*}(s) - v^*(s)| &= \min_{a\in\mathcal{A}_s} \left[r(s,a) + \gamma P(\cdot|s,a)^T V^{\pi_1,\pi_2^*} \right] - \min_{a\in\mathcal{A}_s} \left[r(s,a) + \gamma P(\cdot|s,a)^T V^* \right] \\
&\le \max_{a\in\mathcal{A}_s} \left[\gamma P(\cdot|s,a)^T V^{\pi_1,\pi_2^*} - \gamma P(\cdot|s,a)^T V^* \right] \\
&\le \gamma \|v^{\pi_1,\pi_2^*} - v^*\|_\infty.
\end{aligned}$$

Therefore, we have

$$\|v^{\pi_1,\pi_2^*} - v^*\|_\infty \le \gamma \|v^{\pi_1,\pi_2^*} - v^*\|_\infty + 2\zeta,$$

which indicates that

$$\|v^{\pi_1,\pi_2^*} - v^*\|_\infty \le \frac{2\zeta}{1-\gamma}.$$

The first inequality of (14) is verified. □

Proof of Theorem 3. We define

$$\mathcal{T}V(s) = \begin{cases} \max_{a\in\mathcal{A}_s} r(s,a) + \gamma P_{s,a}^T V, & \forall s \in \mathcal{S}_1, \\ \min_{a\in\mathcal{A}_s} r(s,a) + \gamma P_{s,a}^T V, & \forall s \in \mathcal{S}_2, \end{cases}$$

$$\hat{\mathcal{T}}^{(t)}V(s) = \begin{cases} \max_{a\in\mathcal{A}_s} r(s,a) + \gamma \phi(s,a)^T \Phi_{\mathcal{K}}^{-1} \hat{P}_{\mathcal{K}}^{(t)} V, & \forall s \in \mathcal{S}_1, \\ \min_{a\in\mathcal{A}_s} r(s,a) + \gamma \phi(s,a)^T \Phi_{\mathcal{K}}^{-1} \hat{P}_{\mathcal{K}}^{(t)} V, & \forall s \in \mathcal{S}_2, \end{cases}$$

where $\hat{P}^{(t)} \in \mathbb{R}^{K \times |\mathcal{S}|}$ $(1 \le t \le R)$ is the approximate transition probability matrix obtained by sampling at t-th iterations:

$$\hat{P}^{(t)}_{(s,a),s'} = \frac{1}{T}\sum_{i=1}^{T} 1_{[s_i^{(t)}=s']}, \qquad \forall (s,a) \in \mathcal{K}, \quad s_1^{(t)}, \cdots, s_T^{(t)} \sim P(\cdot|s,a).$$

Then our algorithm can be written as

$$V_{w^{(t)}} \leftarrow \hat{\mathcal{T}}^{(t)} \hat{V}_{w^{(t-1)}}, \quad \forall 1 \le l \le R,$$

where $\hat{V}_{w^{(t-1)}} = \Pi_{[0,1/(1-\gamma)]} V_{w^{(t-1)}}$.

We define the following event as $\mathcal{E}_t$:

$$\|(\hat{P}^{(t)}_{\mathcal{K}} - P_{\mathcal{K}})\hat{V}_{w^{t-1}}\|_\infty \le \epsilon_1 := c \cdot \frac{1}{1-\gamma}\sqrt{\frac{\log(KR/\delta)}{T}}.$$

According to Hoeffding inequality for both state-action pairs in $\mathcal{K}$ and applying their union bound, the event $\mathcal{E}_t$ holds with probability at least $1 - \delta/R$. Also $\mathcal{E}_t$ indicates that for any $s \in \mathcal{S}, a \in \mathcal{A}_s$, we have

$$\left|\gamma\phi(s,a)^T \Phi_{\mathcal{K}}^{-1} \hat{P}^{(t)}_{\mathcal{K}} \hat{V}_{w^{(t-1)}} - \gamma P_{s,a}^T \hat{V}_{w^{(t-1)}}\right| \le \left|\phi(s,a)^T \Phi_{\mathcal{K}}^{-1}(\hat{P}^{(t)}_{\mathcal{K}} - P_{\mathcal{K}})\hat{V}_{w^{(t-1)}}\right| \le L\epsilon_1, \quad (15)$$

where we let $P_{s,a} = P(\cdot|s,a)$ and the last inequality comes from the assumption $\|\phi(s,a)\|_1 \le 1$ and $\|\Phi_{\mathcal{K}}^{-1}\|_\infty \le L$. Noting that for any $s \in \mathcal{S}$,

$$|[\hat{\mathcal{T}}^{(t)}\hat{V}_{w^{(t-1)}}](s) - [\mathcal{T}\hat{V}_{w^{(t-1)}}](s)| \le \max_{a \in \mathcal{A}_s}\left|\gamma\phi(s,a)^T \Phi_{\mathcal{K}}^{-1}\hat{P}^{(t)}_{\mathcal{K}}\hat{V}_{w^{(t-1)}} - \gamma P_{s,a}^T \hat{V}_{w^{(t-1)}}\right|,$$

and $\mathcal{T}v^* = v^*$, we have

$$\begin{aligned}
\|V_{w^{(t)}} - v^*\|_\infty &\le \|\hat{\mathcal{T}}^{(t)}\hat{V}_{w^{(t-1)}} - \mathcal{T}\hat{V}_{w^{(t-1)}}\|_\infty + \|\mathcal{T}\hat{V}_{w^{(t-1)}} - \mathcal{T}v^*\|_\infty \\
&\le L\epsilon_1 + \|\mathcal{T}\hat{V}_{w^{(t-1)}} - \mathcal{T}v^*\|_\infty \\
&\le L\epsilon_1 + \gamma\|\hat{V}_{w^{(t-1)}} - v^*\|_\infty,
\end{aligned}$$

where in the last inequality we use the contraction property of $\mathcal{T}$.

Therefore, when $\mathcal{E}^{(1)}, \cdots, \mathcal{E}^{(R)}$ all hold, we get

$$\begin{aligned}
\|V_{w^{(R)}} - V^*\|_\infty &\le L\epsilon_1 + \gamma\|\hat{V}_{w^{(R-1)}} - v^*\|_\infty \\
&\le L\epsilon_1 + \gamma\|V_{w^{(R-1)}} - v^*\|_\infty \\
&\le \gamma \cdot L\epsilon_1 + L\epsilon_1 + \gamma^2\|\hat{V}_{w^{(R-2)}} - v^*\|_\infty \\
&\le \cdots \\
&\le \frac{\gamma}{1-\gamma} \cdot L\epsilon_1 + \frac{\gamma^R}{1-\gamma},
\end{aligned}$$

where we use the fact $\|V_{w^{(0)}} - V^*\|_\infty \le 1/(1-\gamma)$ in the last inequality. Furthermore, according to $Q_{w^{(R)}}(s,a) = r(s,a) + \gamma\phi(s,a)^T\Phi_{\mathcal{K}}^{-1}\hat{P}^{(R)}_{\mathcal{K}}\hat{V}_{w^{(R-1)}}$, we have

$$\|Q_{w^{(R)}} - Q^*\|_\infty \le L\epsilon_1 + \gamma\|\hat{V}_{w^{(R-1)}} - v^*\|_\infty \le \frac{\gamma \cdot L\epsilon_1}{1-\gamma} + \frac{\gamma^R}{1-\gamma},$$

when $\mathcal{E}^{(R)}$ holds, similar to the derivation in (15). Finally if we suppose $\overline{\pi}^1_{w^{(R)}}, \overline{\pi}^2_{w^{(R)}}$ are optimal counterstrategies against $\pi^2_{w^{(R)}}, \pi^1_{w^{(R)}}$, using Lemma 1 we get

$$\left\|V^{\pi^1_{w^{(R)}}, \overline{\pi}^2_{w^{(R)}}} - V^*\right\|_\infty \le \frac{2\gamma \cdot L\epsilon_1}{(1-\gamma)^2} + \frac{2\gamma^R}{(1-\gamma)^2}$$

$$\left\|V^{\overline{\pi}^1_{w^{(R)}}, \pi^2_{w^{(R)}}} - V^*\right\|_\infty \le \frac{2\gamma \cdot L\epsilon_1}{(1-\gamma)^2} + \frac{2\gamma^R}{(1-\gamma)^2}$$

if events $\mathcal{E}_1, \cdots, \mathcal{E}_R$ all hold. Since every $\mathcal{E}_t$ holds with probability at least $1 - \delta/R$, the probability when all these events hold is at least $1 - \delta$.

Hence if we choose

$$R = \frac{1}{1-\gamma} \log \frac{1}{\epsilon(1-\gamma)}, \quad \epsilon_1 = \mathcal{O}\left(\frac{\epsilon(1-\gamma)^2}{L}\right), \quad T = \mathcal{O}\left(\frac{\log(KR/\delta)}{\epsilon_1^2(1-\gamma)^2}\right) = \mathcal{O}\left(\frac{L^2 \cdot \log(KR/\delta)}{\epsilon^2(1-\gamma)^6}\right),$$

we can obtain a strategy $\pi^{(R)}$ such that

$$\left\|V^{\pi^1_{w^{(R)}}, \overline{\pi}^2_{w^{(R)}}} - V^*\right\|_\infty \leq \epsilon, \quad \left\|V^{\overline{\pi}^1_{w^{(R)}}, \pi^2_{w^{(R)}}} - V^*\right\|_\infty \leq \epsilon$$

with probability at least $1 - \delta$. And the number of samples required is

$$T \cdot R \cdot k = \tilde{\mathcal{O}}\left(\frac{kL^2}{\epsilon^2(1-\gamma)^7}\right).$$

□

C Variance Reduced Q Learning Algorithm

In this section, we will present the whole algorithm described in Section 3.

Algorithm 2 One-side Parametric Q-Learning with Variance Reduction for 2-TBSG

Input: A two-TBSG $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with feature map ϕ, where $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2, \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$
Input: $\epsilon, \delta \in (0, 1)$
Initialize:

$$R' \leftarrow \Theta(\log 1/(\epsilon(1-\gamma))), \quad R \leftarrow \Theta(R'/(1-\gamma)), \quad \theta^{(0,0)} \leftarrow \{\mathbf{0}\} \in \mathbb{R}^K$$
$$m \leftarrow \Theta(L^2 \log(R'RK\delta^{-1})/(\epsilon^2(1-\gamma)^4)), \quad m_1 \leftarrow \Theta(L^2 \log(R'RK\delta^{-1})/((1-\gamma)^2)),$$
$$\epsilon_1 \leftarrow \Theta[L/(1-\gamma) \cdot \sqrt{\log(RR'K\delta^{-1})/m}],$$
$$\epsilon^{(i)} \leftarrow \epsilon_1 + \Theta[L \cdot 2^{-i}/(1-\gamma)\sqrt{\log(RR'K\delta^{-1})/m_1}], \quad \forall 0 \leq i \leq R$$

for $i = 0, 1, \cdots, R'$ **do**
  for $k \in [K]$ **do**
    Generate state samples $x_k^{(1)}, \cdots, x_k^{(m)} \in \mathcal{S}$ from $P(\cdot|s_k, a_k)$ for $(s_k, a_k) \in \mathcal{K}$.
    Let $M^{(i,0)}(k) \leftarrow \frac{1}{m}\sum_{l=1}^m V_{\theta^{(i,0)}}(x_k^{(l)})$
  end for
  for $j = 1, \cdots, R$ **do**
    for $k \in [K]$ **do**
      Generate state samples $x_k^{(1)}, \cdots, x_k^{(m_1)} \in \mathcal{S}$ from $P(\cdot|s_k, a_k)$ for $(s_k, a_k) \in \mathcal{K}$.
      Let $M^{(i,j)}(k) \leftarrow \frac{1}{m_1}\sum_{l=1}^{m_1}\left(V_{\theta^{(i,j-1)}}(x_k^{(l)}) - V_{\theta^{(i,0)}}(x_k^{(l)})\right) + M^{(i,0)}$
    end for
    $w^{(i,j)} \leftarrow \Phi_{\mathcal{K}}^{-1} M^{(i,j)}$
    $\overline{w}^{(i,j)}(k) \leftarrow w^{(i,j)}(k) - \epsilon^{(i)}, \quad \forall k \in [K]$
    $\theta^{(i,j)} \leftarrow \theta^{(i,j-1)} \cup \{\overline{w}^{(i,j)}\}$
  end for
  $\theta^{(i+1,0)} \leftarrow \theta^{(i,R)}$
end for
Output: $\theta^{(R',R)}$

D Construction of Non-negative Feature

In this section, we discuss how to construct nonnegative features such that the Assumption 1 holds.

Theorem 4. *Suppose the transition model P can be embedded into ϕ. If there exists a set $\mathcal{K}$ of state-action pair such that $|\mathcal{K}| = K$ and $\Phi_{\mathcal{K}}$ is nonsingular, then we can construct dimension-$(k+1)$ nonnegative features $\phi'(s, a)$ for every state action pairs (s, a) such that P can also be embedded into ϕ', and Assumption 1 holds for a state-action pair set $\mathcal{K}'$ and some constant L.*

Algorithm 3 Two-side Parametric Q-Learning with Variance Reduction

1: **Input:** A two-TBSG $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with feature map ϕ, where $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2, \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$
2: **Input:** $\epsilon, \delta \in (0, 1)$
3: Construct $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P, 1 - r, \gamma)$ where objectives of two players are switched
4: Solve $\theta^{(R',R)}$ using Algorithm 2 with input $\mathcal{M}$ and ϵ, δ
5: Solve $\eta^{(R',R)}$ using Algorithm 2 with input $\mathcal{M}'$ and ϵ, δ
6: Construct a strategy $\pi : \mathcal{S} \to \mathcal{A}$:

$$\pi(s) = \begin{cases} \pi_{\theta^{(R',R)}}(s), & \forall s \in \mathcal{S}_1, \\ \pi'_{\eta^{(R',R)}}(s), & \forall s \in \mathcal{S}_2. \end{cases} \tag{16}$$

7: **Output:** π

We first present a lemma which is useful in proving Theorem 4.

Lemma 2. *If A is a non-singular matrix and $k \in [K]$, and $A'(l)$ is the following matrix*

$$A'(l)_{i,j} = \begin{cases} A_{i,j} + l, & \text{if } i = k; \\ A_{i,j}, & \text{if } i \neq k. \end{cases}$$

Then for any $N > 0$, there exists $l > N$ such that $A'(l)$ is non-singular.

Proof. This lemma directly follows from the fact that $\det(A'(l))$ is a linear function of l. □

Proof of Theorem 4. We construct a feature map $\mathcal{L}$ from $\phi'(s, a)$ to $\phi(s, a)$ in the following way.

$$\begin{aligned} &\mathcal{L}\left([\phi'_1(s,a), \phi'_2(s,a), \cdots, \phi'_K(s,a), \phi'_{K+1}(s,a)]\right) \\ &= [\phi'_1(s,a) - \phi'_{K+1}(s,a), \phi'_2(s,a) - \phi'_{K+1}(s,a), \cdots, \phi'_K(s,a) - \phi'_{K+1}(s,a)] \\ &= [\phi_1(s,a), \phi_2(s,a), \cdots, \phi_K(s,a)], \end{aligned} \tag{17}$$

which means $\phi_k(s, a) = \phi'_k(s, a) - \phi'_{K+1}(s, a)$ for any $k \in [K]$. Adopting this feature map, P can be embedded into ϕ'. Hence we only need to construct nonnegative features ϕ' satisfying both (17) and Assumption 1.

For any (s, a) there exists $N(s, a)$ such that for any $\phi'_{K+1}(s, a) \geq N(s, a)$, we have $\phi_k(s, a) + \phi'_{K+1}(s, a) \geq 0$ for any $k \in [K]$. We choose arbitrarily a state-action pair (s', a') not in $\mathcal{K}$, and let $\mathcal{K}'$ to be the union of $\mathcal{K}$ and $\{(s', a')\}$. If we choose $\phi'_{K+1}(s', a') > \max\{0, N(s', a')\}$ and $\phi'_k(s', a') = \phi_k(s', a') + \phi'_{K+1}(s', a')$, we will have

$$\det \begin{bmatrix} \phi_1(s_1,a_1) & \cdots & \phi_K(s_1,a_1) & 0 \\ \phi_1(s_2,a_2) & \cdots & \phi_K(s_2,a_2) & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \phi_1(s_K,a_K) & \cdots & \phi_K(s_K,a_K) & 0 \\ \phi'_1(s',a') & \cdots & \phi'_K(s',a') & \phi'_{K+1}(s',a') \end{bmatrix} \neq 0$$

since $\Phi_{\mathcal{K}}$ is nonsingular, where $\mathcal{K} = \{(s_1, a_1), \cdots, (s_K, a_K)\}$. Next for $k \in [K]$, we iteratively choose $\phi'_{K+1}(s_k, a_k) \geq N(s_k, a_k)$ and add it to the k-th row of the feature matrix $\Phi'_{\mathcal{K}'}$ such that the matrix is still nonsingular. (According to Lemma 2, such $\phi'_{K+1}(s_k, a_k)$ exists.) After these K operations, we have

$$\det \Phi'_{\mathcal{K}'} = \det \begin{bmatrix} \phi'_1(s_1,a_1) & \cdots & \phi'_K(s_1,a_1) & \phi'_{K+1}(s_1,a_1) \\ \phi'_1(s_2,a_2) & \cdots & \phi'_K(s_2,a_2) & \phi'_{K+1}(s_2,a_2) \\ \vdots & \ddots & \vdots & \vdots \\ \phi'_1(s_K,a_K) & \cdots & \phi'_K(s_K,a_K) & \phi'_{K+1}(s_K,a_K) \\ \phi'_1(s',a') & \cdots & \phi'_K(s',a') & \phi'_{K+1}(s',a') \end{bmatrix} \neq 0,$$

where $\phi'_k(s_k, a_k) = \phi_k(s_k, a_k) + \phi'_{K+1}(s_k, a_k)$. This indicates that $\Phi'_{\mathcal{K}'}$ is nonsingular. Next for (s, a) not in $\mathcal{K}'$, we choose $\phi'_{K+1}(s, a) = N(s, a)$ and let $\phi'_k(s, a) = \phi_k(s, a) + \phi'_{K+1}(s, a)$.

Then we have $\phi'(s, a) \geq 0$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\Phi'_{\mathcal{K}'}$ is nonsingular. Finally we normalize all features such that $\|\phi'(s, a)\|_1 \leq 1$ while keeping $\phi'(s, a) \geq 0$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. And Assumption 1 holds for $L = \|(\Phi'_{\mathcal{K}})^{-1}\|_\infty$. □

E Proof of Theorem 1

In this section, we present the formal proof of Theorem 1. In the following, we assume that all features $\phi(s,a)$ are nonnegative, and Assumption 1 holds for all the time.

E.1 Notations

We define the following $\mathcal{T}$-operators and Q functions:

$$[\mathcal{T}V](s) = \begin{cases} \max_{a\in\mathcal{A}} r(s,a) + \gamma P(\cdot|s,a)^T V, & \forall s \in \mathcal{S}_1, \\ \min_{a\in\mathcal{A}} r(s,a) + \gamma P(\cdot|s,a)^T V, & \forall s \in \mathcal{S}_2, \end{cases}$$
$$[\mathcal{T}_\pi V](s) = r(s,\pi(s)) + \gamma P(\cdot|s,\pi(s))^T V,$$

$$Q_{\theta^{(i,j)}}(s,a) = r(s,a) + \gamma \phi(s,a)^T \overline{w}^{(i,j)},$$
$$\overline{Q}_{\theta^{(i,j)}}(s,a) = r(s,a) + \gamma P(\cdot|s,a)^T V_{\theta^{(i,j-1)}}.$$

For these $\mathcal{T}$ and $\mathcal{T}_\pi$ operators, we have the following monotonicity and contraction property:

Proposition 1. *For any value function V, V', if $V \le V'$, we have*

$$\mathcal{T}V \le \mathcal{T}V', \quad \mathcal{T}_\pi V \le \mathcal{T}_\pi V'$$
$$\|\mathcal{T}V - \mathcal{T}V'\|_\infty \le \gamma\|V - V'\|_\infty, \quad \|\mathcal{T}_\pi V - \mathcal{T}_\pi V'\|_\infty \le \gamma\|V - V'\|_\infty$$

for any strategy π. Furthermore, v^* and V^π are unique fixed points of $\mathcal{T}$ and $\mathcal{T}_\pi$, respectively, and

$$\lim_{t\to\infty} \mathcal{T}^t V = v^*, \quad \lim_{t\to\infty} \mathcal{T}_\pi^t V = v^\pi.$$

Next we use the following events to describe properties of our algorithm.

- Let $\mathcal{G}^{(i)}$ to be the event
$$0 \le V_{\theta^{(i,0)}}(s) \le [\mathcal{T}V_{\theta^{(i,0)}}](s) \le v^*(s),$$
$$V_{\theta^{(i,0)}}(s) \le \left[\mathcal{T}_{\pi_{\theta^{(i,0)}}} V_{\theta^{(i,0)}}\right](s),$$
$$v^*(s) - V_{\theta^{(i,0)}}(s) \le \frac{2^{-i}}{1-\gamma};$$
- $\mathcal{E}^{(i,0)}$ to be the event of
$$\|w^{(i,0)} - \Phi_\mathcal{K}^{-1} P_\mathcal{K} V_{\theta^{(i,0)}}\|_\infty \le \epsilon_1,$$
where $w^{(i,0)} = \Phi_\mathcal{K}^{-1} M^{(i,0)}$;
- $\mathcal{E}^{(i,j)}$ to be the event of
$$\|w^{(i,j)} - w^{(i,0)} - \Phi_\mathcal{K}^{-1} P_\mathcal{K}(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}})\|_\infty \le \Theta\left[\frac{L\cdot 2^{-i}}{1-\gamma}\sqrt{\frac{\log(RR'K\delta^{-1})}{m_1}}\right].$$

E.2 Preserving the Monotonicity

We first present several lemmas to establish some properties of $\mathcal{T}$ and $\mathcal{T}_{\pi_{\theta^{(i,j)}}}$ on $V_{\theta^{(i,j)}}$.

Lemma 3. *Suppose $\mathcal{G}^{(i)}, \mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,j)}$ holds. We have*

$$0 \le V_{\theta^{(i,j')}}(s) \le [\mathcal{T}V_{\theta^{(i,j')}}](s) \le v^*(s),$$
$$V_{\theta^{(i,j')}}(s) \le \left[\mathcal{T}_{\pi_{\theta^{(i,j')}}} V_{\theta^{(i,j')}}\right](s),$$

for $\forall 0 \le j' \le j$.

Proof. We prove this result by induction. When $j' = 0$, these conditions already hold according to the event $\mathcal{G}^{(i)}$. Now assuming these conditions hold for $j' - 1 \geq 0$, we consider the case of j'.

According to the construction of $V_{\theta(i,j')}$,

$$\begin{aligned} V_{\theta(i,j')}(s) &= \max\left\{V_{\theta(i,j'-1)}(s), \max_{a \in \mathcal{A}_s} Q_{\theta(i,j)}(s,a)\right\} \geq V_{\theta(i,j'-1)}(s), \quad \forall s \in \mathcal{S}_1, \\ V_{\theta(i,j')}(s) &= \max\left\{V_{\theta(i,j'-1)}(s), \min_{a \in \mathcal{A}_s} Q_{\theta(i,j)}(s,a)\right\} \geq V_{\theta(i,j'-1)}(s), \quad \forall s \in \mathcal{S}_2, \end{aligned} \tag{18}$$

Hence for any $s \in \mathcal{S}$, there are two cases:

1. $V_{\theta(i,j')}(s) = V_{\theta(i,j'-1)}(s)$. Then $\pi_{\theta(i,j')}(s) = \pi_{\theta(i,j'-1)}(s)$;
2. (a) $V_{\theta(i,j')}(s) = \max_a Q_{\theta(i,j')}(s,a)$ and $\pi_{\theta(i,j')} = \arg\max_{a \in \mathcal{A}_s} Q_{\theta(i,j')}(s,a)$ if $s \in \mathcal{S}_1$;
   (b) $V_{\theta(i,j')}(s) = \min_a Q_{\theta(i,j')}(s,a)$ and $\pi_{\theta(i,j')} = \arg\min_{a \in \mathcal{A}_s} Q_{\theta(i,j')}(s,a)$ if $s \in \mathcal{S}_2$.

In the first case, according to induction results and the monotonicity of $\mathcal{T}, \mathcal{T}_\pi$, we have

$$V_{\theta(i,j')}(s) = V_{\theta(i,j'-1)}(s) \leq [\mathcal{T} V_{\theta(i,j'-1)}](s) \leq [\mathcal{T} V_{\theta(i,j')}](s)$$

$$V_{\theta(i,j')}(s) = V_{\theta(i,j'-1)}(s) \leq \left[\mathcal{T}_{\pi_{\theta(i,j'-1)}} V_{\theta(i,j'-1)}\right](s) = \left[\mathcal{T}_{\pi_{\theta(i,j')}} V_{\theta(i,j'-1)}\right](s) \leq \left[\mathcal{T}_{\pi_{\theta(i,j')}} V_{\theta(i,j')}\right](s)$$

In the second case, according to the event $\mathcal{E}^{(i,0)}, \mathcal{E}^{(i,j')}$, we obtain

$$\begin{aligned} &\|w^{(i,j')} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V_{\theta(i,j'-1)}\|_\infty \\ &\leq \|w^{(i,j')} - w^{(i,0)} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} (V_{\theta(i,j'-1)} - V_{\theta(i,0)})\|_\infty + \|w^{(i,0)} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V_{\theta(i,0)}\|_\infty \\ &\leq \Theta\left[\frac{L \cdot 2^{-i}}{1-\gamma}\sqrt{\frac{\log(RR'K\delta^{-1})}{m_1}}\right] + \epsilon_1 \\ &= \epsilon^{(i)}, \end{aligned}$$

which indicates that

$$\overline{w}^{(i,j')} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V_{\theta(i,j'-1)} = w^{(i,j')} - \epsilon^{(i)} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V_{\theta(i,j'-1)} \leq 0.$$

Since all features are nonnegative, we have the following inequality for $\forall a \in \mathcal{A}_s$.

$$\begin{aligned} Q_{\theta(i,j')}(s,a) &= r(s,a) + \gamma\phi(s,a)^T \overline{w}^{(i,j')} \leq r(s,a) + \gamma\phi(s,a)^T \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V_{\theta(i,j'-1)} \\ &= r(s,a) + \gamma P(\cdot|s,a)^T V_{\theta(i,j'-1)} = \overline{Q}_{\theta(i,j')}(s,a). \end{aligned}$$

Therefore, we have

$$\begin{aligned} V_{\theta(i,j')}(s) &= \max_{a \in \mathcal{A}_s} Q_{\theta(i,j')}(s,a) \leq \max_{a \in \mathcal{A}_s} \overline{Q}_{\theta(i,j')}(s,a) = [\mathcal{T} V_{\theta(i,j'-1)}](s), \quad \text{if } s \in \mathcal{S}_1; \\ V_{\theta(i,j')}(s) &= \min_{a \in \mathcal{A}_s} Q_{\theta(i,j')}(s,a) \leq \min_{a \in \mathcal{A}_s} \overline{Q}_{\theta(i,j')}(s,a) = [\mathcal{T} V_{\theta(i,j'-1)}](s), \quad \text{if } s \in \mathcal{S}_2; \\ V_{\theta(i,j')}(s) &= Q_{\theta(i,j')}(s, \pi_{\theta(i,j')}(s)) \leq \overline{Q}_{\theta(i,j')}(s, \pi_{\theta(i,j')}(s)) = \left[\mathcal{T}_{\pi_{\theta(i,j')}} V_{\theta(i,j'-1)}\right](s). \end{aligned}$$

Noticing $V_{\theta(i,j'-1)} \leq V_{\theta(i,j')}$ and the monotonicity of $\mathcal{T}$ and $\mathcal{T}_{\pi_{\theta(i,j'-1)}}$, we have

$$\begin{aligned} V_{\theta(i,j')}(s) &\leq [\mathcal{T} V_{\theta(i,j'-1)}](s) \leq [\mathcal{T} V_{\theta(i,j')}](s), \\ V_{\theta(i,j')}(s) &\leq \left[\mathcal{T}_{\pi_{\theta(i,j'-1)}} V_{\theta(i,j')}\right](s). \end{aligned}$$

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

$$\begin{aligned} V_{\theta(i,j')}(s) &\leq [\mathcal{T} V_{\theta(i,j')}](s), \\ V_{\theta(i,j')}(s) &\leq \left[\mathcal{T}_{\pi_{\theta(i,j'-1)}} V_{\theta(i,j')}\right](s). \end{aligned}$$

Again according to the monotonicity of $\mathcal{T}$, we have

$$V_{\theta(i,j')}(s) \leq [\mathcal{T} V_{\theta(i,j')}](s) \leq \left[\mathcal{T}^2 V_{\theta(i,j')}\right](s) \leq \cdots \leq v^*(s).$$

This completes the induction. $\square$

Next, we will exhibit an approximate contraction property of our algorithm.

Lemma 4. *If $\mathcal{G}^{(i)}, \mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,j)}$ holds, then for $1 \leq j' \leq j$ we have*

$$v^*(s) - V_{\theta^{(i,j')}}(s) \leq \max_{a \in \mathcal{A}_s} \left[\gamma P(\cdot|s,a)^T(v^* - V_{\theta^{(i,j'-1)}})\right] + 2\epsilon^{(i)}, \quad \forall s \in \mathcal{S}.$$

Proof. According to Algorithm 2,

$$\overline{w}^{(i,j')} = w^{(i,j')} - \epsilon^{(i)} \geq \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}^T V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i)}.$$

Using $V_{\theta^{(i,j'-1)}} \leq V_{\theta^{(i,j')}} \leq v^*$ in Lemma 3, we have for $\forall s \in \mathcal{S}_1$,

$$\begin{aligned}
v^*(s) - V_{\theta^{(i,j')}}(s) &\leq v^*(s) - \max_{a \in \mathcal{A}_s}\left[r(s,a) + \gamma\phi(s,a)^T\overline{w}^{(i,j')}\right] \\
&\leq v^*(s) - \max_{a \in \mathcal{A}_s}\left[r(s,a) + \gamma\phi(s,a)^T(\Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}^T V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i)})\right] \\
&= \max_{a \in \mathcal{A}_s}\left[r(s,a) + \gamma P(\cdot|s,a)^T v^*\right] - \max_{a \in \mathcal{A}_s}\left[r(s,a) + \gamma\phi(s,a)^T(\Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}^T V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i)})\right] \\
&\leq \max_{a \in \mathcal{A}_s}\left[\gamma P(\cdot|s,a)^T v^* - \gamma P(\cdot|s,a)^T V_{\theta^{(i,j'-1)}} + 2\epsilon^{(i)}\|\phi(s,a)\|_1\right] \\
&\leq \max_{a \in \mathcal{A}_s}\left[\gamma P(\cdot|s,a)^T(v^* - V_{\theta^{(i,j'-1)}})\right] + 2\epsilon^{(i)},
\end{aligned}$$

where the first inequality is due to $V_{\theta^{(i,j')}}(s) \geq \max_{a \in \mathcal{A}_s}\left[r(s,a) + \gamma\phi(s,a)^T\overline{w}^{(i,j')}\right]$, and the third equality is due to the property of v^*. And for $s \in \mathcal{S}_2$, similarly we have

$$\begin{aligned}
v^*(s) - V_{\theta^{(i,j')}}(s) &\leq v^*(s) - \min_{a \in \mathcal{A}_s}[r(s,a) + \gamma\phi(s,a)^T\overline{w}^{(i,j')}] \\
&\leq v^*(s) - \min_{a \in \mathcal{A}_s}\left[r(s,a) + \gamma\phi(s,a)^T(\Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}^T V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i)})\right] \\
&= \min_{a \in \mathcal{A}_s}\left[r(s,a) + P(\cdot|s,a)^T v^*\right] - \min_{a \in \mathcal{A}_s}\left[r(s,a) + \gamma\phi(s,a)^T(\Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}^T V_{\theta^{(i,j'-1)}} - 2\epsilon^{(i)})\right] \\
&\leq \max_{a \in \mathcal{A}_s}\left[\gamma P(\cdot|s,a)^T v^* - \gamma P(\cdot|s,a)^T V_{\theta^{(i,j'-1)}} + 2\epsilon^{(i)}\|\phi(s,a)\|_1\right] \\
&\leq \max_{a \in \mathcal{A}_s}\left[\gamma P(\cdot|s,a)^T(v^* - V_{\theta^{(i,j'-1)}})\right] + 2\epsilon^{(i)}.
\end{aligned}$$

The proof is completed. □

E.3 Analysis of the Confidence Bounds

In this subsection, we analyze the confidence bound of sampling.

Lemma 5. *For $0 \leq i \leq R'$,*

$$\mathbf{Pr}\left(\mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,R)}|\mathcal{G}^{(i)}\right) \leq 1 - \frac{\delta}{R'}$$

Proof. Conditioned on $\mathcal{G}^{(i)}$, we have

$$\begin{aligned}
0 \leq V_{\theta^{(i,0)}}(s) \leq v^*(s) \leq \frac{1}{1-\gamma}, \\
v^*(s) - V_{\theta^{(i,0)}}(s) \leq \frac{2^{-i}}{1-\gamma}.
\end{aligned}$$

For any $k \in [K]$ and $\delta \in (0,1)$, according to Hoeffding inequality, with probability at least $1-\delta$,

$$|M^{(i,0)}(k) - P(\cdot|s_k,a_k)V_{\theta^{(i,0)}}| \leq c_1 \cdot \max_{s \in \mathcal{S}}|V_{\theta^{(i,0)}}| \cdot \sqrt{\frac{\log[\delta^{-1}]}{m}} \leq c_1 \cdot \frac{1}{1-\gamma}\sqrt{\frac{\log[\delta^{-1}]}{m}}$$

holds for some constant c_1. If we switch δ into $\delta/(RR'K)$, then we obtain

$$|M^{(i,0)}(k) - P(\cdot|s_k,a_k)V_{\theta^{(i,0)}}| \leq \epsilon_1/L$$

holds with probability at least $1-\delta/(RR'K)$. Next using the fact $\|\Phi_{\mathcal{K}}\|_\infty \le L$ and apply the union bound for all $k \in [K]$, we have

$$\|w^{(i,0)} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V_{\theta^{(i,0)}}\|_\infty = \|\Phi_{\mathcal{K}}^{-1}(M^{(i,0)} - P_{\mathcal{K}} V_{\theta^{(i,0)}})\|_\infty \le \epsilon_1$$

holds with probability at least $1-\delta/(RR')$. This indicates that $\mathcal{E}^{(i,0)}$ holds with probability at least $1-\delta/(RR')$.

As for $\mathcal{E}^{(i,1)}$, since $M^{(i,1)} = M^{(i,0)}$ and $w^{(i,1)} = w^{(i,0)}$, the event $\mathcal{E}^{(i,1)}$ holds with probability 1.

For $2 \le j \le R$, again using the Hoeffding inequality and the event $\mathcal{G}^{(i)}$, we have

$$\begin{aligned}
&|M^{(i,j)}(k) - M^{(i,0)}(k) - P(\cdot|s_k, a_k)^T \left(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}}\right)| \\
=& \left|\frac{1}{m_1}\sum_{l=1}^{m_1}\left(V_{\theta^{(i,j-1)}}(x_k^{(l)}) - V_{\theta^{(i,0)}}(x_k^{(l)})\right) - P(\cdot|s_k, a_k)^T \left(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}}\right)\right| \\
\le& c_1 \max_{s\in\mathcal{S}} |V_{\theta^{(i,j-1)}}(s) - V_{\theta^{(i,0)}}(s)| \cdot \sqrt{\frac{\log(\delta^{-1})}{m_1}} \\
\le& c_1 \max_{s\in\mathcal{S}} |v^*(s) - V_{\theta^{(i,0)}}(s)| \cdot \sqrt{\frac{\log(\delta^{-1})}{m_1}} \\
\le& c_1 \cdot \frac{2^{-i}}{1-\gamma}\sqrt{\frac{\log(\delta^{-1})}{m_1}}
\end{aligned}$$

with probability at least $1-\delta$. Since $w^{(i,j)} = \Phi_{\mathcal{K}}^{-1} M^{(i,j)}, w^{(i,0)} = \Phi_{\mathcal{K}}^{-1} M^{(i,0)}$, we switch δ into $\delta/(RR'K)$ and apply the union bound for all $k \in [K]$ to obtain that

$$\begin{aligned}
&\|w^{(i,j)} - w^{(i,0)} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}})\|_\infty \\
=& \|\Phi_{\mathcal{K}}^{-1}\left(M^{(i,j)} - M^{(i,0)} - P_{\mathcal{K}}(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}})\right)\|_\infty \\
\le& L \cdot \|M^{(i,j)} - M^{(i,0)} - P_{\mathcal{K}}(V_{\theta^{(i,j-1)}} - V_{\theta^{(i,0)}})\|_\infty \\
\le& \Theta\left[\frac{L\cdot 2^{-i}}{1-\gamma}\sqrt{\frac{\log(RR'K\delta^{-1})}{m_1}}\right]
\end{aligned}$$

holds with probability at least $1-\delta/(RR')$. So is the probability of $\mathcal{E}^{(i,j)}$ conditioned on $\mathcal{G}^{(i)}$. Applying the union bound for all $\mathcal{E}^{(i,0)}, \mathcal{E}^{(i,1)}, \mathcal{E}^{(i,2)} \cdots, \mathcal{E}^{(i,R)}$, we have

$$\mathbf{Pr}\left(\mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,R)}|\mathcal{G}^{(i)}\right) \le 1 - \frac{\delta}{R'},$$

which completes the proof. □

E.4 Analysis of the Error in the Next Iteration

Lemma 6. *For* $\forall 1 \le i \le R'$, *we have*

$$\mathbf{Pr}(\mathcal{G}^{(i+1)}, \mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,R)}|\mathcal{G}^{(i)}) \ge 1 - \frac{\delta}{R'}.$$

Proof. Conditioned on $\mathcal{G}^{(i)}$, suppose $\mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,R)}$ all hold. Using Lemma 4 for R times, there exists a constant C such that

$$\begin{aligned}
\|v^* - V_{\theta^{(i,R)}}\|_\infty &\leq \max_{s\in\mathcal{S}, a\in\mathcal{A}} \left[\gamma P(\cdot|s,a)^T (v^* - V_{\theta^{(i,R-1)}})\right] + 2\epsilon^{(i)} \\
&\leq \gamma \|v^* - V_{\theta^{(i,R-1)}}\|_\infty + 2\epsilon^{(i)} \\
&\leq \cdots \\
&\leq \gamma^R \|v^* - V_{\theta^{(i,0)}}\|_\infty + 2\sum_{j'=0}^{R-1} \gamma^{j'} \epsilon^{(i)} \\
&\leq \frac{\gamma^R}{1-\gamma} + \frac{C}{1-\gamma}\left[\frac{2^{-i}L}{1-\gamma}\sqrt{\frac{\log(RR'K\delta^{-1})}{m_1}} + \frac{L}{1-\gamma}\sqrt{\frac{\log(RR'K\delta^{-1})}{m}}\right],
\end{aligned}$$

where the last inequality is due to events $\mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,R)}$ and $\|v^* - V_{\theta^{(i,0)}}\|_\infty \leq 1/(1-\gamma)$ according to $\mathcal{G}^{(i)}$. Hence if we choose

$$\begin{aligned}
R &= C_R \cdot \frac{\log(1/(\epsilon(1-\gamma)))}{1-\gamma}, \quad m = C_1 \cdot \frac{L^2 \log(R'RK\delta^{-1})}{\epsilon^2(1-\gamma)^4}, \\
m_1 &= C_2 \cdot \frac{L^2 \log(R'RK\delta^{-1})}{(1-\gamma)^2}, \quad \epsilon \leq \frac{2^{-i}}{1-\gamma},
\end{aligned}$$

where C_R, C_1, C_2 are constant numbers, we will have

$$\|v^* - V_{\theta^{(i,R)}}\|_\infty \leq \frac{2^{-i-1}}{1-\gamma}.$$

Furthermore, since $\theta^{(i+1,0)} = \theta^{(i,R)}$, the following inequalities

$$\begin{aligned}
&0 \leq V_{\theta^{(i+1,0)}}(s) \leq [\mathcal{T}V_{\theta^{(i+1,0)}}](s) \leq v^*(s), \\
&V_{\theta^{(i+1,0)}}(s) \leq \left[\mathcal{T}_{\pi_{\theta^{(i+1,0)}}} V_{\theta^{(i+1,0)}}\right](s) \\
&v^*(s) - V_{\theta^{(i+1,0)}}(s) \leq \frac{2^{-i-1}}{1-\gamma}
\end{aligned}$$

hold according to Lemma 3. Hence the event $\mathcal{G}^{(i+1)}$ holds when $\mathcal{G}^{(i)}, \mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,R)}$ all hold. Therefore, according to Lemma 5 we have proved that

$$\mathbf{Pr}(\mathcal{G}^{(i+1)}, \mathcal{E}^{(i,0)}, \cdots, \mathcal{E}^{(i,R)} | \mathcal{G}^{(i)}) \geq 1 - \frac{\delta}{R'}.$$

□

E.5 Analysis of Approximation from Two Sides

Lemma 7. *With probability at least $1-\delta$, $\mathcal{G}^{(0)}, \mathcal{G}^{(i)}, \mathcal{E}^{(i-1,j)}$ hold for $\forall 1 \leq i \leq R', 0 \leq j \leq R$ and $V_{\theta^{(R',R)}}$ is an ϵ-optimal value.*

Proof. First of all, according to the initialization,

$$V_{\theta^{(0,0)}}(s) = \begin{cases} \max_{a\in\mathcal{A}_s} r(s,a) \geq 0, & \text{if } s \in \mathcal{S}_1, \\ \min_{a\in\mathcal{A}_s} r(s,a) \geq 0, & \text{if } s \in \mathcal{S}_2. \end{cases}$$

This indicates that

$$[\mathcal{T}V_{\theta^{(0,0)}}](s) = \begin{cases} \max_{a\in\mathcal{A}_s} r(s,a) + \gamma P(\cdot|s,a)^T V_{\theta^{(0,0)}} \geq \max_{a\in\mathcal{A}_s} r(s,a) = V_{\theta^{(0,0)}}(s), & \text{if } s \in \mathcal{S}_1, \\ \min_{a\in\mathcal{A}_s} r(s,a) + \gamma P(\cdot|s,a)^T V_{\theta^{(0,0)}} \geq \min_{a\in\mathcal{A}_s} r(s,a) = V_{\theta^{(0,0)}}(s), & \text{if } s \in \mathcal{S}_2. \end{cases}$$

and

$$\left[\mathcal{T}_{\pi_{\theta^{(0,0)}}} V_{\theta^{(0,0)}}\right](s) = r(s, \pi_{\theta^{(0,0)}}) + \gamma P(\cdot|s, \pi_{\theta^{(0,0)}}) V_{\theta^{(0,0)}}(s) \geq r(s, \pi_{\theta^{(0,0)}}) = V_{\theta^{(0,0)}}(s).$$

According to the monotonicity of $\mathcal{T}$,

$$0 \leq V_{\theta^{(0,0)}} \leq \mathcal{T}V_{\theta^{(0,0)}} \leq \cdots \leq v^* \leq \frac{1}{1-\gamma}.$$

Hence $\mathcal{G}^{(0)}$ always holds.

Next based on Lemma 6, we have

$$\begin{aligned}
&\mathbf{Pr}(\mathcal{G}^{(i)}, \mathcal{E}^{(i-1,j)}, \forall 1 \leq i \leq R', 0 \leq j \leq R | \mathcal{G}^{(0)}) \\
= &\mathbf{Pr}(\mathcal{G}^{(i)}, \mathcal{E}^{(i-1,j)}, \forall 1 \leq i \leq R'-1, 0 \leq j \leq R | \mathcal{G}^{(0)}) \mathbf{Pr}(\mathcal{G}^{(R')}, \mathcal{E}^{(R'-1,0)}, \cdots, \mathcal{E}^{(R'-1,R)} | \mathcal{G}^{(R'-1)}) \\
\geq &\mathbf{Pr}(\mathcal{G}^{(i)}, \mathcal{E}^{(i-1,j)}, \forall 1 \leq i \leq R'-1, 0 \leq j \leq R | \mathcal{G}^{(0)}) \cdot (1-\delta/R') \\
\geq &\mathbf{Pr}(\mathcal{G}^{(i)}, \mathcal{E}^{(i-1,j)}, \forall 1 \leq i \leq R'-2, 0 \leq j \leq R | \mathcal{G}^{(0)}) \cdot (1-\delta/R')^2 \\
\geq &\cdots \\
\geq &(1-\delta/R')^{R'} \\
\geq &1-\delta.
\end{aligned}$$

If $\mathcal{G}^{(R')}$ holds, then we obtain

$$\|v^* - V_{\theta^{(R,R')}}\|_\infty \leq \frac{2^{-R'}}{1-\gamma}.$$

Hence when choosing $R' = \Theta[\log(\epsilon^{-1}(1-\gamma)^{-1})]$, we have

$$\|v^* - V_{\theta^{(R,R')}}\|_\infty \leq \epsilon,$$

which indicates that $V_{\theta^{(R,R')}}$ is an ϵ-optimal value. Therefore, the event that $V_{\theta^{(R,R')}}$ is an ϵ-optimal value, together with $\mathcal{G}^{(i)}, \mathcal{E}^{(i,j)}$, happens with probability at least $1-\delta$. $\square$

Next, we provide some notations and lemmas for the 2-TBSG $\mathcal{M}'$. Suppose the equilibrium value function of $\mathcal{M}'$ is v'. For $\eta = \{z^{(h)}\}_{h=1}^Z$, we define

$$V'_\eta(s) = \begin{cases} \max_{h\in[Z]} \min_{a\in\mathbb{A}_s} \left(r'(s,a) + \gamma\phi(s,a)^T z^{(h)}\right), & \forall s \in \mathcal{S}_1, \\ \max_{h\in[Z]} \max_{a\in\mathbb{A}_s} \left(r'(s,a) + \gamma\phi(s,a)^T z^{(h)}\right), & \forall s \in \mathcal{S}_2, \end{cases}$$

$$\pi'_\eta(s) = \begin{cases} \max_{h\in[Z]} \left[\arg\min_{a\in\mathcal{A}_s} \left(r'(s,a) + \gamma\phi(s,a)^T z^{(h)}\right)\right], & \forall s \in \mathcal{S}_1, \\ \max_{h\in[Z]} \left[\arg\max_{a\in\mathcal{A}_s} \left(r'(s,a) + \gamma\phi(s,a)^T z^{(h)}\right)\right], & \forall s \in \mathcal{S}_2, \end{cases}$$

$$W_\eta(s) = \frac{1}{1-\gamma} - V'_\eta(s).$$

To describe the connection between values of $\mathcal{M}$ and $\mathcal{M}'$, we introduce the following lemma, whose proof is obvious.

Lemma 8. $v'(s) = 1/(1-\gamma) - v(s)$. *Any equilibrium strategy for $\mathcal{M}'$ is also an equilibrium strategy for $\mathcal{M}$. And if $\|V'_\eta - v'\| \leq \epsilon$, then W_η is an ϵ-optimal value.*

Next we define events $\mathcal{G}^{(0)}, \mathcal{G}_1^{(i)}, \mathcal{E}_1^{(i-1,j)}$ for $1 \leq i \leq R', 0 \leq j \leq R$ similar to the case of $\mathcal{M}$.

- Let $\mathcal{G}_1^{(i)}$ to be the event

$$0 \leq V'_{\eta^{(i,0)}}(s) \leq \left[\mathcal{T}'V'_{\eta^{(i,0)}}\right](s) \leq v'(s),$$

$$V'_{\eta^{(i,0)}}(s) \leq \left[\mathcal{T}'_{\pi'_{\eta^{(i,0)}}} V'_{\eta^{(i,0)}}\right](s),$$

$$v'(s) - V'_{\eta^{(i,0)}}(s) \leq \frac{2^{-i}}{1-\gamma},$$

  where $\mathcal{T}'$ and $\mathcal{T}'_{\pi'}$ is defined as

$$[\mathcal{T}'V'](s) = \begin{cases} \min_{a\in\mathcal{A}_s}[r'(s,a) + \gamma P(\cdot|s,a)^T V], & \forall s \in \mathcal{S}_1, \\ \max_{a\in\mathcal{A}_s}[r'(s,a) + \gamma P(\cdot|s,a)^T V], & \forall s \in \mathcal{S}_2. \end{cases} \tag{19}$$

$$[\mathcal{T}'_{\pi'}V'](s) = r'(s,\pi'(s)) + \gamma P(\cdot|s,\pi'(s))^T V', \quad \forall s \in \mathcal{S}.$$

This event is equivalent to the event

$$\frac{1}{1-\gamma} \geq W_{\eta^{(i,0)}}(s) \geq \left[\mathcal{T} W_{\eta^{(i,0)}}\right](s) \geq v^*(s),$$
$$W_{\eta^{(i,0)}}(s) \geq \left[\mathcal{T}_{\pi'_{\eta^{(i,0)}}} W_{\eta^{(i,0)}}\right](s),$$
$$v^*(s) - W_{\eta^{(i,0)}}(s) \geq \frac{2^{-i}}{1-\gamma};$$

- $\mathcal{E}_1^{(i,0)}$ to be the event of
$$\|z^{(i,0)} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}} V'_{\eta^{(i,0)}}\|_\infty \leq \epsilon_1;$$
- $\mathcal{E}_1^{(i,j)}$ to be the event of
$$\|z^{(i,j)} - z^{(i,0)} - \Phi_{\mathcal{K}}^{-1} P_{\mathcal{K}}(V'_{\eta^{(i,j-1)}} - V'_{\eta^{(i,0)}})\|_\infty \leq \Theta\left[\frac{L \cdot 2^{-i}}{1-\gamma}\sqrt{\frac{\log(RR'K\delta^{-1})}{m_1}}\right].$$

We present two following lemmas, which can be viewed as counterparts for W of Lemma 3 and Lemma 7.

Lemma 9. *Suppose $\mathcal{G}_1^{(i)}, \mathcal{E}_1^{(i,0)}, \cdots, \mathcal{E}_1^{(i,j)}$ holds. We have*

$$1/(1-\gamma) \geq W_{\theta^{(i,j')}}(s) \geq [\mathcal{T} W_{\theta^{(i,j')}}](s) \geq v^*(s),$$
$$W_{\theta^{(i,j')}}(s) \geq \left[\mathcal{T}_{\pi'_{\eta^{(i,j')}}} W_{\theta^{(i,j')}}\right](s),$$

for $\forall 0 \leq j' \leq j$.

Proof. The proof is similar to Lemma 3. □

Lemma 10. *With at least probability $1-\delta$, Algorithm 2 for $\mathcal{M}'$ will output $\eta^{(R,R')}$ which satisfies $\|V'_{\eta^{(R,R')}} - v'\| \leq \epsilon$, together with events $\mathcal{G}_1^{(0)}, \mathcal{G}_1^{(i)}, \mathcal{E}_1^{(i-1,j)}$ for $1 \leq i \leq R', 0 \leq j \leq R$.*

Proof. The proof is similar to Lemma 7. □

Our next lemma indicates that if V_θ and W_η are both ϵ-optimal values, then the strategy obtained from Algorithm 3 is an ϵ-optimal strategy.

Lemma 11. *If $V_{\theta^{(R',R)}}$ and $W_{\eta^{(R',R)}}$ are both ϵ-optimal values, where $\theta^{(R',R)}, \eta^{(R',R)}$ are parameters obtained from Algorithm 2 with inputs $\mathcal{M}$ and $\mathcal{M}'$, respectively. and $\mathcal{G}^{(0)}, \mathcal{G}^{(i)}, \mathcal{E}^{(i-1,j)}$, $\mathcal{G}_1^{(0)}, \mathcal{G}_1^{(i)}, \mathcal{E}_1^{(i-1,j)}$ all hold for $\forall 1 \leq i \leq R', 0 \leq j \leq R$, then the strategy $\pi = (\pi_1, \pi_2)$ output from Algorithm 3 is an ϵ-optimal strategy.*

Proof. We define the following operators mapping from value functions to value functions.

$$[\mathcal{T}_{\max,\pi_2} V](s) = \begin{cases} \max_{a \in \mathcal{A}_s}[r(s,a) + \gamma P(\cdot|s,a)^T]V, & \forall s \in \mathcal{S}_1, \\ r(s,\pi_2(s)) + \gamma P(\cdot|s,\pi_2(s))^T V, & \forall s \in \mathcal{S}_2, \end{cases}$$
$$[\mathcal{T}_{\pi_1,\min} V](s) = \begin{cases} r(s,\pi_1(s)) + \gamma P(\cdot|s,\pi_1(s))^T V, & \forall s \in \mathcal{S}_1, \\ \min_{a \in \mathcal{A}_s}[r(s,a) + \gamma P(\cdot|s,a)^T]V, & \forall s \in \mathcal{S}_2. \end{cases}$$

Then $\mathcal{T}_{\max,\pi_2}, \mathcal{T}_{\pi_1,\min}$ are both monotonic and contracting operators with contraction factor γ, and it is easy to see that $V^{\overline{\pi}_1,\pi_2}, V^{\pi_1,\overline{\pi}_2}$ are fixed points of $\mathcal{T}_1, \mathcal{T}_2$, respectively, where $\overline{\pi}_1, \overline{\pi}_2$ are optimal counterstrategies against π_2, π_1.

We next prove that $V^{\overline{\pi}_1,\pi_2}$ satisfies

$$v^* \leq V^{\overline{\pi}_1,\pi_2} \leq W_{\eta^{(R',R)}}. \tag{20}$$

According to Lemma 9, for any $s \in \mathcal{S}_1$, we have

$$\left[\mathcal{T}_{\max,\pi_2} W_{\eta^{(R',R)}}\right](s) = \max_{a\in\mathcal{A}_s} r(s,a) + \gamma P(\cdot|s,a)^T W_{\eta^{(R',R)}} = \left[\mathcal{T} W_{\eta^{(R',R)}}\right](s) \le W_{\eta^{(R',R)}}(s),$$

and for any $s \in \mathcal{S}_2$, we have

$$\begin{aligned}\left[\mathcal{T}_{\max,\pi_2} W_{\eta^{(R',R)}}\right](s) &= r(s,\pi_2(s)) + \gamma P(\cdot|s,\pi_2(s))^T W_{\eta^{(R',R)}} \\ &= \left[\mathcal{T}_{\pi'_{\eta^{(R',R)}}} W_{\eta^{(R',R)}}\right](s) \le W_{\eta^{(R',R)}}(s).\end{aligned}$$

These inequalities indicates that $\mathcal{T}_{\max,\pi_2} W_{\eta^{(R',R)}} \le W_{\eta^{(R',R)}}$. Hence according to the monotonicity of $\mathcal{T}_{\max,\pi_2}$, we have $V^{\overline{\pi}_1,\pi_2} \le W_\eta^{(R',R)}(s)$. Moreover, since $\overline{\pi}_1$ is an optimal counterstrategy against π_2, we have $v^* \le V^{\overline{\pi}_1,\pi_2}$. The inequalities (20) has been proved.

Next noticing that $\|W_{\eta^{(R',R)}} - v^*\| \le \epsilon$, we have

$$\|V^{\overline{\pi}_1,\pi_2} - v^*\| \le \epsilon.$$

Similarly we have $\|V^{\pi_1,\overline{\pi}_2} - v^*\| \le \epsilon$ considering the operator $\mathcal{T}_{\pi_1,\min}$. These two inequalities together indicate that π is an ϵ-optimal strategy. $\square$

E.6 Proof of Theorem 1

Proof of Theorem 1. According to Lemma 7 and 10, the event that $V_{\theta^{(R,R')}}, W_{\eta^{(R,R')}}$ are both ϵ-optimal values, together with events $\mathcal{G}^{(0)}, \mathcal{G}^{(i)}, \mathcal{E}^{(i,j)}, \mathcal{G}_1^{(0)}, \mathcal{G}_1^{(i)}, \mathcal{E}_1^{(i,j)}$ for $1 \le i \le R', 0 \le j \le R$, happen with probability at least $1 - 2\delta$. Hence according to Lemma 11, the output π of Algorithm 3 is ϵ-optimal strategy with probability at least $1 - 2\delta$. The total samples used in our algorithm is

$$2(R'RKm_1 + R'Km) = \tilde{\mathcal{O}}\left(\frac{KL^2}{\epsilon^2(1-\gamma)^4}\right)$$

samples. $\square$

F Proof of Theorem 2

We first present a proposition indicating that an approximate optimal strategy of $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is also an approximate optimal strategy of $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P', r, \gamma)$.

Proposition 2. *Suppose $P, \tilde{P}$ are two transition models such that*

$$\|P(\cdot|s,a) - P'(\cdot|s,a)\|_{TV} \le \xi, \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Then for two 2-TBSGs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma), \mathcal{M}' = (\mathcal{S}, \mathcal{A}, P', r, \gamma)$, if π is an ϵ-optimal strategy of $\mathcal{M}$, π is also an $\left(\frac{2\xi}{(1-\gamma)^2} + 2\epsilon\right)$-optimal strategy of $\mathcal{M}'$.

Proof. Suppose $\pi = (\pi_1, \pi_2), \overline{\pi}_1, \overline{\pi}_2$ are optimal counterstrategies against π_2, π_1 in $\mathcal{M}$, and $\overline{\pi}'_1, \overline{\pi}'_2$ are optimal counterstrategy against π_2, π_1 in $\mathcal{M}'$. We also assume that V, U are value functions of $\mathcal{M}$ and $\mathcal{M}'$, respectively.

According to the TV condition, for any strategy π we have

$$\begin{aligned}\|V^\pi - U^\pi\|_\infty &= \|(I-\gamma P_\pi)^{-1} r_\pi - (I-\gamma P'_\pi)^{-1} r_\pi\|_\infty \\ &= \|(I-\gamma P'_\pi)^{-1}(\gamma P_\pi - \gamma P'_\pi)(I-\gamma P_\pi)^{-1} r_\pi\|_\infty \\ &\le \|(I-\gamma P'_\pi)^{-1}\|_\infty \|\gamma P_\pi - \gamma P'_\pi\|_\infty \|(I-\gamma P_\pi)^{-1}\|_\infty \|r_\pi\|_\infty \\ &\le \frac{1}{1-\gamma\|P'_\pi\|_\infty} \cdot \|\gamma P_\pi - \gamma P'_\pi\|_\infty \cdot \frac{1}{1-\gamma\|P_\pi\|_\infty} \|r_\pi\|_\infty \\ &\le \frac{\xi}{(1-\gamma)^2},\end{aligned}$$

where the last inequality follows the facts

$$\|P_\pi\|_\infty = \|P'_\pi\|_\infty = 1, \quad \|r_\pi\|_\infty \le 1, \quad \|P_\pi - P'_\pi\|_\infty = \max_{s\in\mathcal{S}} \|P(\cdot|s,\pi(s)) - P'(\cdot|s,\pi(s))\|_{TV} \le \xi.$$

Next, since $\overline{\pi}_2, \overline{\pi}'_2$ are optimal counterstrategies against π_1 in $\mathcal{M}$ and $\mathcal{M}'$, we have

$$V^{\pi_1,\overline{\pi}_2} \le V^{\pi_1,\overline{\pi}'_2}, \quad U^{\pi_1,\overline{\pi}_2} \ge U^{\pi_1,\overline{\pi}'_2}.$$

Hence for any $s \in \mathcal{S}$,

$$-\frac{\xi}{(1-\gamma)^2} \le V^{\pi_1,\overline{\pi}_2}(s) - U^{\pi_1,\overline{\pi}_2} \le V^{\pi_1,\overline{\pi}_2}(s) - U^{\pi_1,\overline{\pi}'_2}(s) \le V^{\pi_1,\overline{\pi}'_2}(s) - U^{\pi_1,\overline{\pi}'_2}(s) \le \frac{\xi}{(1-\gamma)^2},$$

which indicates that

$$\left\|V^{\pi_1,\overline{\pi}_2} - U^{\pi_1,\overline{\pi}'_2}\right\|_\infty \le \frac{\xi}{(1-\gamma)^2}.$$

Similarly, we have

$$\left\|V^{\overline{\pi}_1,\pi_2} - U^{\overline{\pi}'_1,\pi_2}\right\|_\infty \le \frac{\xi}{(1-\gamma)^2}.$$

Moreover, since π is an ϵ-optimal strategy of $\mathcal{M}$, we have

$$\|V^{\pi_1,\overline{\pi}_2} - V^{\overline{\pi}_1,\pi_2}\|_\infty \le \|V^{\pi_1,\overline{\pi}_2} - v^*\| + \|V^{\overline{\pi}_1,\pi_2} - v^*\|_\infty \le 2\epsilon,$$

where v^* is the equilibrium value of $\mathcal{M}'$. This inequality, together with above two inequalities, indicates that

$$\left\|U^{\pi_1,\overline{\pi}'_2} - U^{\overline{\pi}'_1,\pi_2}\right\| \le \frac{2\xi}{(1-\gamma)^2} + 2\epsilon.$$

Next noting that

$$U^{\pi_1,\overline{\pi}'_2} \le u^* \le U^{\overline{\pi}'_1,\pi_2},$$

where u^* is the equilibrium value of $\mathcal{M}'$, we have

$$\left\|U^{\pi_1,\overline{\pi}'_2} - u^*\right\|_\infty \le \frac{2\xi}{(1-\gamma)^2} + 2\epsilon, \quad \left\|U^{\overline{\pi}'_1,\pi_2} - u^*\right\|_\infty \le \frac{2\xi}{(1-\gamma)^2} + 2\epsilon.$$

These two inequalities indicate that π is an $\left(\frac{2\xi}{(1-\gamma)^2} + 2\epsilon\right)$-optimal strategy of $\mathcal{M}'$. □

Proof of Theorem 2. Since Algorithm 1, 2 and 3 only sample from $P(\cdot|s,a)$ for $(s,a) \in \mathcal{K}$, and P and P' agree on $\mathcal{K}$, the results of these algorithms executing on $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ are same as the results of these algorithms executing on $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P', r, \gamma)$. According to Theorem 3 and Theorem 1, with probability at least $1-\delta$, Algorithm 1 outputs $w^{(R)}$ such that $\pi_{w^{(R)}}$ is an ϵ-optimal strategy of $\mathcal{M}$, and with probability at least $1-2\delta$, Algorithm 3 outputs an ϵ-optimal strategy π of $\mathcal{M}$. Therefore, according to Proposition 2, $\pi_{w^{(R)}}$ and π are $\left(\frac{2\xi}{(1-\gamma)^2} + 2\epsilon\right)$-optimal strategies of $\mathcal{M}'$. The proof is completed. □