Gap-Dependent Bounds for Two-Player Markov Games

Zehao Dou
Yale University

Zhuoran Yang
Princeton University

Zhaoran Wang
Northwestern University

Simon S.Du
University of Washington

Abstract

As one of the most popular methods in the field of reinforcement learning, Q-learning has received increasing attention. Recently, there have been more theoretical works on the regret bound of algorithms that belong to the Q-learning class in different settings. In this paper, we analyze the cumulative regret when conducting Nash Q-learning algorithm on 2-player turn-based stochastic Markov games (2-TBSG), and propose the first gap dependent logarithmic upper bounds in the episodic tabular setting. This bound matches the lower bound only up to a horizon term. Furthermore, we extend the conclusion to the discounted game setting with infinite horizon and propose a similar gap dependent logarithmic regret bound. In addition, under the linear MDP assumption, we obtain another logarithmic regret for 2-TBSG, in both centralized and independent settings.

1 INTRODUCTION

Recently, designing an effective and efficient algorithm to obtain a near-optimal strategy in sequential decision-making tasks has attracted an increasing interest in the field of reinforcement learning (RL) (Sutton and Barto, 1988). By estimating the optimal state-action value function (a.k.a Q-function), Q-learning method (Watkins and Dayan, 1992) is one of the most popular classes of algorithms. In each iteration of Q-learning algorithms, the agent greedily chooses the action with the largest Q value and achieves its reward and the next state by interacting with the underlying RL environment. At the same time, the algorithm keeps updating the Q-values by using Bellman Equation. In comparison, the model-based methods attempt to reveal the structure of the environment, which lead to more memory and worse time efficiency. These advantages of Q-learning make it play an important role in a wide range of RL problems (Mnih et al., 2013, 2015).

In this paper, we study a more complicated scenario, two-player turn-based stochastic Markov game, which is a special case of multi-agent reinforcement learning (MARL). In this setting, two players, known as the max-player and the min-player, interact with each other one by one and optimize their individual rewards. The max-player's goal is to maximize the cumulative reward while the min-player attempts to minimize it. In order to measure the quality of the two players' policies, we study the **total regret**, where the players learn a policy tuple (π_k, μ_k) for a sequence of episodes $k = 1, 2, \ldots, K$, and suffer a total regret, which is the total sub-optimality of the policies $\pi_1, \pi_2, \ldots, \pi_K$. From the regret minimization perspective, many related works (Kearns and Singh, 1998; Sm, 2003; Azar et al., 2013; Jin et al., 2018; Dong et al., 2019; Liu and Su, 2020; Bai et al., 2020) have provided a $\sqrt{K}$-type upper bound of total regret in various settings where K is the number of episodes. Although these $\sqrt{K}$-type upper bounds are easy to understand and match the lower bound, they paint an overly pessimistic worst-case scenario on the Markov decision processes.

Recent works (Yang et al., 2020; Ok et al., 2018; He et al., 2020; Simchowitz and Jamieson, 2019; Xu et al., 2021) limit the MDP with certain structures and propose much tighter upper bound or provide new perspectives. One line of works assume the existence of minimal sub-optimality gap, and establish the $C \log K$-type of total regret upper bound, where C is an instance-dependent constant associated with the minimal sub-optimality gap, named $\text{gap}_{\min}$:

$$\text{gap}_{\min} := \min\{\text{gap}(s,a) = V^*(s) - Q^*(s,a) > 0\}.$$

Here, V^* and Q^* denote the value function and Q-function for an optimal policy π^*.

Another line of works (Xie et al., 2020; Jin et al., 2019; He et al., 2020) assume a certain structure of reward function and probability transition kernel, such as the linear function approximation. For instance, Jin et al.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

(2019) studies the episodic MDPs with linear MDP assumptions, which means that both the transition probability function and reward function can be represented as a linear function of a given feature mapping. He et al. (2020) combines the two ideas and provides a gap-dependent logarithmic regret bound with linear function approximation.

However, all the provable logarithmic regret bounds are under the single agent setting where the sole player attempts to achieve the highest cumulative rewards by interacting with the underlying environment. For the two-player settings or multi-agent settings, the gap-dependent logarithmic regret bound still remains absent because the sub-optimality gaps are not non-negative and the concept of minimal sub-optimality gap no longer makes sense in these settings. In this paper, we overcome these difficulties and propose a gap-dependent logarithmic cumulative regret bound of 2-TBSG in all the tabular, discounted and linear function expression settings for the first time. Next, we introduce the three main contributions of this paper, which are listed below.

- In two-player episodic turn-based general sum stochastic games (2-TBSG) with finite horizon (Jia et al., 2019; Shapley, 1953), we propose a new concept named minimal positive sub-optimality gap $\text{gap}^+_{\min}$. Based on that, we provide a gap-dependent logarithmic total regret bound $\mathcal{O}\left(\frac{H^6SA\log(SAT)}{\text{gap}^+_{\min}}\right)$ when using the Nash Q-learning algorithm.
- Based on the result above, we further extend it to the discounted 2-TBSG with infinite horizon, and obtain a total regret upper bound $\mathcal{O}\left(\frac{SA}{\text{gap}^+_{\min}(1-\gamma)^5\log(1/\gamma)}\cdot\log\frac{SAT}{\text{gap}^+_{\min}(1-\gamma)}\right)$, which is gap dependent and logarithmic on T.
- We analyze the finite-horizon 2-TBSG with linear function expression. Under the linear MDP assumption, we propose the 2-TBSG version of LSVI-UCB algorithm (Jin et al., 2019) and provide a provable $\widetilde{\mathcal{O}}\left(\frac{d^3H^5\log(16dK^3H^3)}{\text{gap}^+_{\min}}\right)$ total regret upper bound which is also gap dependent and logarithmic on K, in both centralized and independent settings.

Technically, we are using a new set of algorithms when solving the pure Nash Equilibrium of 2-TBSG, which provably exists, and it makes our proof novel and more difficult. Compared with Xie et al. (2020) that establishes $\sqrt{T}$-regret, to achieve the gap-dependent logarithmic regret, we utilize a different regret decomposition method links the regret to a sum of gap terms, which are further bounded via a peeling argument. See Sections B.1 and B.3 for details. Meanwhile, compared with He et al. (2020); Du et al. (2019), we face new difficulties since the game setting we are analyzing has a max-player as well as a min-player. Therefore, we need to propose a new technique to control the influence of the opponent, which results in developing both upper confidence bound (UCB) and lower confidence bound (LCB) at the same time. To our best knowledge, our result establish the first logarithmic regret bounds for zero-sum Markov games under both the tabular and linear settings.

2 RELATED WORKS

Tabular and Infinite-horizon MDP There is a long list of results focusing on the regret or sample complexity on tabular episodic MDPs and discounted MDP with infinite horizon. They can be recognized as model-free methods or model-based methods, which are two different types of methodology. Model-based methods (Jaksch et al., 2010; Dann et al., 2017; Osband et al., 2016) explicitly estimate the transition probability function while the model-free methods (Jin et al., 2018; Strehl et al., 2006) do not. One line of works (Sidford et al., 2018; Lattimore and Hutter, 2012; Ghavamzadeh et al., 2011; Wainwright, 2019; Azar et al., 2012; Koenig and Simmons, 1993) assume the existence of a simulator (also called a generative model) where the agent can freely query any state-action pair to the underlying environment and return the reward as well as the next state. In the episodic setting without simulators, Jin et al. (2018) achieves a $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ regret bound for a model-free algorithm and Azar et al. (2017) proposes a UCB-VI algorithm with Bernstein style bonus with achieves a $\widetilde{\mathcal{O}}(\sqrt{H^2SAT})$ regret bound for a model-based algorithm. Both of the two upper bounds nearly attain the minimax lower bound $\Omega(\sqrt{H^2SAT})$ (Jaksch et al., 2010; Jin et al., 2018; Osband and Roy, 2016). Recently, Zhang et al. (2020); Jin et al. (2018) provide a $\sqrt{T}$-type regret bound for Q-learning algorithms (which is a widely used model-free algorithm) where T is the number of episodes.

Another line of works focus on providing $\log T$-type regret bound based on instance-dependent quantities. Ok et al. (2018) shows us that $\log T$ is unavoidable as a lower bound. Tewari and Bartlett (2007) proposes an OLP algorithm for average-reward MDP and achieves an asymptotic logarithmic regret $\mathcal{O}(C\log T)$ where the constant C is instance-dependent. Jaksch et al. (2010) provides a UCRL2 algorithm and provides a non-asymptotic regret bound $\mathcal{O}(D^2S^2A\log T/\text{gap}_{\min})$ where D is the diameter and $\text{gap}_{\min}$ is the minimal sub-optimality gap. A recent work (Wang et al., 2019)

proves a $\mathcal{O}(SAH^6 \log(SAT)/\text{gap}_{min})$ regret bound for the model-free optimistic Q-learning algorithm.

Linear Function Approximation A recent line of works (Jin et al., 2018; Wang et al., 2019; Yang and Wang, 2020; Jia et al., 2019; Zanette et al., 2020; Du et al., 2019, 2020; Zhou et al., 2020; Weisz et al., 2020) solve MDP with linear function approximations and propose a regret bound. After parameterizing the Q-function with feature mapping under the linear MDP assumption, Jin et al. (2019) proposes LSVI-UCB algorithm with $\widetilde{\mathcal{O}}(\sqrt{d^3H^3T})$ total regret bound. Zanette et al. (2020) improves the bound to $\widetilde{\mathcal{O}}(\sqrt{d^2H^2T})$ by introducing a global planning oracle. Later, a new linear mixed model assumption is proposed and a number of works (Jia et al., 2019, 2020) introduce the UCLR-VTR algorithm to solve MDP under the new assumption, and provide $\widetilde{\mathcal{O}}(\sqrt{d^2H^3T})$ regret bound. In the discounted setting, Zhou et al. (2020) provides a $\widetilde{\mathcal{O}}(d\sqrt{T}/(1-\gamma)^2)$ upper bound where γ is the discount ratio.

3 PRELIMINARIES AND NOTATIONS

In this section, we introduce some important concepts, notations and background knowledge.

Setting of Two-player Turn-based Stochastic Games In two-player turn-based games (2-TBSG), only one player takes his action at each step. Denote the two players as P_1, P_2, which are the max-player and the min-player respectively. We partition the whole action space as $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, where $\mathcal{A}_i$ is the state space of player P_i. Since the stochastic game is episodic under our setting, we denote $2H$ as the number of steps in one episode. At each step $h \in [2H]$, when h is an odd number, it's the max-player P_1's turn to observe the current state s_h and take an action $a_h \in \mathcal{A}_1$, and then we receive the reward $r_h(s_h, a_h)$. Similarly, when h is an even number, the min-player P_2 observes the current state s_h and takes the action $a_h \in \mathcal{A}_2$, and then they receive the reward $r_h(s_h, a_h)$. After taking the action, the system makes transition to a new state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$. For each action $a \in \mathcal{A}$, denote $I(a) \in \{1, 2\}$ as it indicates of the current player to play, so that $a \in \mathcal{A}_{I(a)}$.

General Notations We denote the tabular episodic Markov Game as $\text{MG}(2H, \mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, \mathbb{P}, r)$, where $2H$ is the number of steps in each episode, $\mathcal{S}$ is the set of states, and $(\mathcal{A}_1, \mathcal{A}_2)$ are the action sets of the max-player and the min-player respectively. Since we analyze the two-player turned base games in this paper, the odd number steps are the max-player's turn and the even number steps are the min-player's turn. $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[2H]}$ is the collection of transition matrices and $\mathbb{P}_h(\cdot|s, a)$ outputs the probability distribution over states when a is the action taken at step h after state s. $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic reward function at step h.

Markov policy, Q-function and Value function We denote (π, μ) as the Markov policy of the max-player and the min-player respectively. Max-player's policy π is a collection of H mappings $\{\pi_h : \mathcal{S} \to \mathcal{A}_1\}_{h=1,3,\cdots,2H-1}$, which maps the current state to the action. Similarly, min-player's policy μ is a collection of H mappings $\{\mu_h : \mathcal{S} \to \mathcal{A}_2\}_{h=2,4,\cdots,2H}$. We denote $\pi_h(s)$ and $\mu_h(s)$ to represent the following action to take under Markov policy π, μ. Next, we define the well-known value functions and Q-functions. we denote $V_h^{\pi,\mu} : \mathcal{S} \to \mathbb{R}$ as the value function at step h under Markov policy π, μ which calculates the expected cumulative rewards:

$$V_h^{\pi,\mu}(s) := \mathbb{E}_{\pi,\mu}\left[\sum_{h'=h}^{2H} r_{h'}(s_{h'}, a_{h'}) \ : \ s_h = s\right].$$

We also define $Q_h^{\pi,\mu} : \mathcal{S} \times \mathcal{A}$ as the Q-function at step h so that $Q_h^{\pi,\mu}(s, a)$ calculates the cumulative rewards under policy (π, μ), starting from (s, a) at step h:

$$Q_h^{\pi,\mu}(s, a) := \mathbb{E}_{\pi,\mu}\left[\sum_{h'=h}^{2H} r_{h'}(s_{h'}, a_{h'}) \ : \ s_h = s, a_h = a\right].$$

It's worth mentioned that the Markov policy of these two players are both deterministic, which means given a current state, their policies point to a specific action, instead of a probability distribution over all actions. That is because under the two-player turn based game setting, the Nash Equilibria is simply a pure strategy.

For simplicity, we introduce the commonly-used notation of operator $\mathbb{P}_h$ which is $[\mathbb{P}_h(V)](s, a) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}V(s')$ for any value function V. By this definition, we have the Bellman equation as follows:

$$Q_h^{\pi,\mu}(s, a) = (r_h + \mathbb{P}_h V_{h+1}^{\pi,\mu})(s, a)$$

holds for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [2H]$. We define $V_{2H+1}^{\pi,\mu}(s) = 0$ for all $s \in \mathcal{S}_{2H+1}$.

Best response and Nash equilibria Given the Markov policy π of the max-player P_1, their exists a best response for the min-player, which we denote as $\mu^\dagger(\pi)$. It satisfies:

$$V_h^{\pi,\mu^\dagger(\pi)}(s) = \inf_\mu V_h^{\pi,\mu}(s)$$

holds for any $(s, h) \in \mathcal{S} \times [2H]$. For brevity, we denote $V_h^{\pi,\dagger} := V_h^{\pi,\mu^\dagger(\pi)}$. Similarly, we can also denote

$\pi^\dagger(\mu), V_h^{\dagger,\mu}$ as the best response for the max-player and the corresponding value function. Furthermore, it is well known that there exists Markov policies π^*, μ^* which are the best responses of each other, and they satisfy:

$$V_h^{\pi^*,\dagger}(s) = \sup_\pi V_h^{\pi,\dagger}(s), \qquad V_h^{\dagger,\mu^*}(s) = \inf_\mu V_h^{\dagger,\mu}(s)$$

holds for $\forall(s,h) \in \mathcal{S} \times [2H]$. This definition of Nash equilibria is Markov perfect Equilibrium, which depends on the states, following the exact same definition in prior work. They can be treated as a solution to the saddle point optimization in the policy space. Here we use inf and sup instead of min and max because the inf and sup are taken from the policy space, which may be infinite or non-compact in a general case.

Discounted 2-TBSG with Infinite Horizon In such a game, the max-player P_1 and the min-player P_2 take turns to take action. However, different from standard 2-TBSG, there is only one episode in this game and there are infinite steps in this episode.

2-TBSG with Linear Function Expression A Markov Game $\mathrm{MG}(2H, \mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ is defined as linear when the probability transition kernels and the reward functions are linear with respect to a given feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Specifically, for each $h \in [2H]$, there exists an unknown vector $\mu_h \in \mathbb{R}^d$ and unknown measures $\theta_h = (\theta_h^{(1)}, \theta_h^{(2)}, \ldots, \theta_h^{(d)})$ whose degree of freedom is $|S| \times d$, such that for $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \theta_h(s') \rangle \text{ and } r(s,a) = \langle \phi(s,a), \mu_h \rangle.$$

For the complete version of the definition, we put it into Assumption 1 below.

Connection with Linear Saddle Point Problem Linear saddle point problem with linear constraints can be formulated as LP while the Markov games can not. According to Schultz (1992), the Markov game proposed by us can only be formulated as a linear complementarity program (LCP), which means given the policy of one player, it is a linear program for the other player. Therefore, it is not simply a linear saddle point problem. That being said, the Markov game can be formulated as a saddle point problem optimization where the variables are the policies of two players. However, such an optimization problem is infinite dimensional as the set of all possible policies is infinite-dimensional and nonlinear because the state transition involves the policies of both players. Intuitively, to see why the optimization problem is nonlinear, notice that the transition operator of the Markov chain induced by the joint policy (π_1, π_2) is given by $P^{\pi_1,\pi_2}(s'|s) = \sum_{a,b} \pi_1(a|s)\pi_2(b|s)P(s'|s,a,b)$. Here the optimization variables π_1 and π_2 are multiplied together, making the optimization problem nonlinear.

Algorithm 1 Optimistic Nash Q-learning on Two-player Turn-based Stochastic Games

Initialize: Let $\overline{Q}_h(s,a) \leftarrow 2H$, $\underline{Q}_h(s,a) \leftarrow 0$, and $N_h(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
Define: $\alpha_t = \frac{2H+1}{2H+t}$, $\iota \leftarrow \log(SAT^2)$.

for episode $k \in [K]$ **do**
  observe the initial state s_1
  for step $h \in [2H]$ **do**
    Take action $a_h \leftarrow \arg\max_{a' \in \mathcal{A}} \overline{Q}_h(s_h, a')$ if h is an odd number, (i.e. $I(a_h) = 1$),else take action $a_h \leftarrow \arg\min_{a' \in \mathcal{A}} \underline{Q}_h(s_h, a')$. After that, observe the next state s_{h+1}.
    $t = N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$.
    $\beta_t \leftarrow c\sqrt{(2H)^3\iota/t}$.
    $\overline{Q}_h(s_h, a_h) \leftarrow (1-\alpha_t) \cdot \overline{Q}_h(s_h, a_h) + \alpha_t \cdot [r_h(s_h, a_h) + \overline{V}_{h+1}(s_{h+1}) + \beta_t]$.
    $\underline{Q}_h(s_h, a_h) \leftarrow (1-\alpha_t) \cdot \underline{Q}_h(s_h, a_h) + \alpha_t \cdot [r_h(s_h, a_h) + \underline{V}_{h+1}(s_{h+1}) - \beta_t]$.
    $\overline{V}_h(s_h) \leftarrow \overline{Q}_h(s_h, a_h)$, $\underline{V}_h(s_h) \leftarrow \underline{Q}_h(s_h, a_h)$.
  end for
end for

Pure Strategies Can be Optimal for 2-TBSG Given a 2-player turn-based stochastic game with horizon $2H$, optimal pure strategy always exists in the final step $2H$, since the optimal action of that player is the one that leads to the optimal reward. If optimal pure strategy exists for horizons $i+1, \ldots, 2H$, then for the i-th step, the optimal action is simply the one with the optimal expected reward given the optimal pure strategy in the following steps. By using the method of induction, we obtain the existence of pure Nash Equilibrium.

Mathematical Notations Let $f(n), g(n)$ be two positive series, we write $f(n) = \mathcal{O}(g(n))$ or $f(n) \lesssim g(n)$ if there exists a positive constant C such that $f(n) \leqslant C \cdot g(n)$ for all n larger than some $n_0 \in \mathbb{N}$. Similarly, we write $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ if there exists a positive constant C such that $f(n) \geqslant C \cdot g(n)$ for all n larger than some $n_0 \in \mathbb{N}$. If these two conditions hold simultaneously, we write $f(n) \asymp g(n)$ or $f(n) = \Theta(g(n))$. If logarithmic terms should be ignored, then we use the notations $\widetilde{\mathcal{O}}, \widetilde{\Omega}, \widetilde{\Theta}$.

4 REGRET BOUND OF TABULAR 2-TBSG

In 2-TBSG, we use a Nash Q-learning method: Algorithm 1 to obtain a policy tuple sequence (π^k, μ^k) to approximate a Nash equilibrium. In this algorithm,

$\overline{Q}_h(s,a), \underline{Q}_h(s,a)$ are the upper and lower estimation of $Q_h^*(s,a)$ respectively. The max-player chooses action based on the upper estimation $\overline{Q}$ while the min-player chooses action based on the lower estimation $\underline{Q}$. By using the Upper Confidential Bound (UCB) technique, we can prove that

$$\underline{Q}_h(s,a) \leqslant Q_h^*(s,a) \leqslant \overline{Q}_h(s,a)$$

holds with high probability. The total regret of this algorithm is defined as:

$$\text{Regret}(K) = \sum_{k=1}^{K} \left| \left(V_1^* - V_1^{\pi^k,\mu^k}\right)(s_1^k) \right|.$$

Here, V_1^* is the abbreviation of $V_1^{\pi^*,\mu^*}$. In this section, we will focus on upper bounding the expected regret $\mathbb{E}[\text{Regret}(K)]$. Notice that, unlike our related papers, there is an absolute value in our definition above, which is because under the two-player game setting, $\left(V_1^* - V_1^{\pi^k,\mu^k}\right)(s_1^k)$ can be either positive or negative.

Theorem 1 (Main Theorem 1: Logarithmic Regret Bound of Q-learning for Episodic 2-TBSG)**.** *After using Algorithm 1, the expected total regret for episodic two-player turn-based stochastic game (2-TBSG) can be upper bounded by:*

$$\mathbb{E}[\text{Regret}(K)] \leqslant \mathcal{O}\left(\frac{H^6 SA \log(SAT)}{\text{gap}_{\min}^+}\right).$$

Here, the definition of $\text{gap}_{\min}^+$ will be left in the following section. Notice that the theorem is not a straightforward extension of existing works since Algorithm 1 is originally proposed by us and it is the first gap dependent logarithmic regret bound for 2-TBSG. In Xie et al. (2020), the authors have proposed a similar algorithm, but their setting is 2-player zero-sum simultaneous-move Markov Games and prove a $\mathcal{O}(\sqrt{d^3H^3T\iota^2})$ regret bound guarantee. The policies of the two players are mixed strategies, which makes it impossible to define gaps. Next, we provide the proof sketch of Theorem 1 and leave all the proof details to appendix.

4.1 First Step: Split the Total Regret into the Expected Sum of Gaps

In the first step, we split the total regret defined above into several single-step sub-optimality gaps.

Lemma 1.

$$\left(V_1^* - V_1^{\pi^k,\mu^k}\right)(s_1^k) = \mathbb{E}\left[\sum_{h=1}^{2H} \text{gap}_h(s_h^k, a_h^k) \mid \pi^k, \mu^k\right].$$

Here, $\text{gap}_h(s_h^k, a_h^k) := V_h^*(s_h^k) - Q_h^*(s_h^k, a_h^k)$.

It is obvious to find that: when $a_h^k \in \mathcal{A}_1$, which means a_h^k is the max-player's action, then $\text{gap}_h(s_h^k, a_h^k) \geqslant 0$. In contrast, when $a_h^k \in \mathcal{A}_2$, which means a_h^k is the min-player's action, then $\text{gap}_h(s_h^k, a_h^k) \leqslant 0$. This is because the optimal strategy group (π^*, μ^*) stands for both the optimal point for max-player and min-player, though at different directions. Therefore, we introduce a new notation:

$$\begin{aligned}\text{gap}_h^+(s_h, a_h) &:= |\text{gap}_h(s_h, a_h)| \\ &= \begin{cases} V_h^*(s_h) - Q_h^*(s_h, a_h) & \text{if } a_h \in \mathcal{A}_1 \text{ (or } h \text{ is odd)} \\ Q_h^*(s_h, a_h) - V_h^*(s_h) & \text{if } a_h \in \mathcal{A}_2 \text{ (or } h \text{ is even)} \end{cases}.\end{aligned}$$

Combining the two equations above, we obtain that:

$$\begin{aligned}\mathbb{E}[\text{Regret}(K)] &= \sum_{k=1}^{K} \left| \mathbb{E}\left[\sum_{h=1}^{2H} \text{gap}_h(s_h^k, a_h^k) \mid \pi^k, \mu^k\right] \right| \\ &\leqslant \sum_{k=1}^{K} \mathbb{E}\left[\sum_{h=1}^{2H} \text{gap}_h^+(s_h^k, a_h^k) \mid \pi^k, \mu^k\right]. \end{aligned} \tag{1}$$

4.2 Second Step: Concentration Inequalities and the Extension

According to an existing lemma in a related work (Bai and Jin, 2020), we have the following conclusion as its special case in the 2-TBSG setting:

Lemma 2 (Lemma 11 of Bai and Jin (2020))**.** *For any $p \in (0,1]$, we let $\iota = \log(SAT/p)$. Then, with probability at least $1-p$, Algorithm 1 has the following property:*

$$\begin{aligned}&\overline{Q}_h^k(s,a) \geqslant Q_h^*(s,a) \geqslant \underline{Q}_h^k(s,a), \\ &\overline{V}_h^k(s) \geqslant V_h^*(s) \geqslant \underline{V}_h^k(s)\end{aligned} \tag{2}$$

holds for $\forall s \in \mathcal{S}, h \in [2H], a \in \mathcal{A}, k \in [K]$. We call the event above $\varepsilon_{\text{conc}}$ and then $P(\varepsilon_{\text{conc}}) \geqslant 1-p$.

This lemma indicates that $\overline{V}, \overline{Q}$ and $\underline{V}, \underline{Q}$ proposed by Algorithm 1 are exactly the upper bound and lower bound of the optimal point (Nash equilibrium) V^*, Q^*, with high probability. Similar to the episodic setting, we can make the following conclusion:

$$\text{gap}_h^+(s_h^k, a_h^k) \leqslant \overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k) \tag{3}$$

holds for $\forall h \in [2H], k \in [K]$ as long as $\varepsilon_{\text{conc}}$ holds. Then, Equation (3) is transformed to:

$$\begin{aligned}&\text{gap}_h^+(s_h^k, a_h^k) = \text{clip}\left[\text{gap}_h^+(s_h^k, a_h^k) \mid \text{gap}_{\min}^+\right] \\ &\leqslant \text{clip}\left[\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k) \mid \text{gap}_{\min}^+\right].\end{aligned} \tag{4}$$

4.3 Third Step: Peeling

Like Yang et al. (2020), we separate all the gaps $\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k)$ into different intervals and

count them individually. Note that when the gap $\overline{Q}_h^k(s_h^k,a_h^k)-\underline{Q}_h^k(s_h^k,a_h^k)$ belongs to $[0,\mathrm{gap}_{\min})$, then it will be clipped to 0 by Equation (4). For the other gaps, we divide them into N different intervals $[\mathrm{gap}_{\min},2\mathrm{gap}_{\min}),\cdots,[2^{N-1}\mathrm{gap}_{\min},2^N\mathrm{gap}_{\min})$. Here, $N=\lceil\log_2(2H/\mathrm{gap}_{\min})\rceil$. The following lemma tells us the upper bound of the counting number in each interval.

Lemma 3 (Bounded Counting Number of Each Interval)**.** *Under the concentration event $\varepsilon_{\mathrm{conc}}$, for each $n\in[N]$, we denote:*

$$C^{(n)}:=\left|\left\{(k,h):\overline{Q}_h^k(s_h^k,a_h^k)-\underline{Q}_h^k(s_h^k,a_h^k)\in\Lambda_n\right\}\right|.$$

where $\Lambda_n=[2^{n-1}\mathrm{gap}_{\min},2^n\mathrm{gap}_{\min})$. Then, we have the following upper bound:

$$C^{(n)}\leqslant\mathcal{O}\left(\frac{H^6SA\iota}{4^n\mathrm{gap}_{\min}^2}\right)$$

where $\iota=\log(SAT/p)$ is the logarithmic term.

Once we have this lemma proved, we can easily estimate the upper bound of expected cumulative regret.

$$\begin{aligned}
&\mathbb{E}[\mathrm{Regret}(K)]\leqslant\mathbb{E}\left[\sum_{k=1}^K\sum_{h=1}^{2H}\mathrm{gap}_h^+\left(s_h^k,a_h^k\right)\right]\\
&\leqslant\sum_{\varepsilon_{\mathrm{conc}}}\mathbb{P}(\mathrm{traj})\cdot\sum_{k=1}^K\sum_{h=1}^{2H}\mathrm{clip}\left[(\overline{Q}_h^k-\underline{Q}_h^k)(s_h^k,a_h^k)\mid\mathrm{gap}_{\min}\right]\\
&\quad+\sum_{\mathrm{traj}\notin\varepsilon_{\mathrm{conc}}}\mathbb{P}(\mathrm{traj})\cdot 2TH\\
&\leqslant\sum_{n=1}^N 2^n\mathrm{gap}_{\min}C^{(n)}+p\cdot 2TH\\
&\leqslant\sum_{n=1}^N\mathcal{O}\left(\frac{H^6SA\iota}{2^n\mathrm{gap}_{\min}}\right)+p\cdot 2TH\leqslant\mathcal{O}\left(\frac{H^6SA\log(SAT)}{\mathrm{gap}_{\min}}\right).
\end{aligned}$$

In the last step, we let $p=\frac{1}{T}$, and then $\iota=\log(SAT^2)=\mathcal{O}(\log(SAT))$. This leads to our main theorem. However, the proof of Lemma 3 is difficult. We rely on a general lemma about the upper bound of the weighted sum of $(\overline{Q}_h^k-\underline{Q}_h^k)(s_h^k,a_h^k)$.

Lemma 4 (Peeling Argument)**.** *Under the event $\varepsilon_{\mathrm{conc}}$, the following holds for $\forall h\in[2H]$ and a weight sequence $\{w_{k,h}\}_{k\in[K]}$ which satisfies: $0\leqslant w_{k,h}\leqslant w$, $\sum_{k=1}^K w_{k,h}\leqslant C$, it holds that:*

$$\begin{aligned}
&\sum_{k=1}^K w_{k,h}\left(\overline{Q}_h^k-\underline{Q}_h^k\right)(s_h^k,a_h^k)\\
&\quad\leqslant 4ewSAH^2+60c\sqrt{SACewH^5\iota}.
\end{aligned}\tag{5}$$

After proving this lemma. We can make

$$w_{k,h}^n=\mathbb{I}\left[\left(\overline{Q}_h^k-\underline{Q}_h^k\right)(s_h^k,a_h^k)\in\Lambda_n\right]$$

and

$$C_h^{(n)}=\sum_{k=1}^K\mathbb{I}\left[\left(\overline{Q}_h^k-\underline{Q}_h^k\right)(s_h^k,a_h^k)\in\Lambda_n\right],$$

where $\Lambda_n=[2^{n-1}\mathrm{gap}_{\min},2^n\mathrm{gap}_{\min})$. Then, the sequence $\{w_{k,h}^n\}_{k\in[K]}$ is a $\left(1,C_h^{(n)}\right)$-sequence. According to Lemma 4, we know that:

$$\begin{aligned}
2^{n-1}\mathrm{gap}_{\min}C_h^{(n)}&\leqslant\sum_{k=1}^K w_{k,h}^n\left(\overline{Q}_h^k-\underline{Q}_h^k\right)(s_h^k,a_h^k)\\
&\leqslant 4eSAH^2+60c\sqrt{SAC_h^{(n)}eH^5\iota},
\end{aligned}$$

which leads to the fact that:

$$C_h^{(n)}\leqslant\mathcal{O}\left(\frac{H^5SA\iota}{4^n\mathrm{gap}_{\min}^2}\right).\tag{6}$$

Finally, after summing them up:

$$C^{(n)}=\sum_{h=1}^{2H}C_h^{(n)}=\mathcal{O}\left(\frac{H^6SA\iota}{4^n(\mathrm{gap}_{\min}^+)^2}\right),$$

which comes to our conclusion of Lemma 3.

Connection Between the Vanilla Regret and Duality Regret. Here we remark on the connection between the vanilla regret and duality regret.

- In this paper, we analyze the vanilla regret $|V_1^*-V_1^{\pi^k,\mu^k}|$. Since the game setting we are analyzing is 2-player turn-based stochastic game (2-TBSG), and (π^*,μ^*) is a Nash Equilibrium rather than the optimal policy. The vanilla regret actually measures the distance between the two value functions proposed by (π^*,μ^*) and (π^k,μ^k).

- Another type of regret is the duality regret $V_1^{\dagger,\mu^k}-V_1^{\pi^k,\dagger}$. It can measure how close the policy (π^k,μ^k) is to a Nash Equilibrium.

Therefore, it's an important question whether a small vanilla regret implies a small duality regret. (Notice that the inverse would not be true since there might be more than one Nash Equilibria.) To answer this question, we propose the following proposition.

Proposition 1. *If for $\forall s\in\mathcal{S},h\in[2H]$, the vanilla regret*

$$|V_h^*(s)-V_h^{\pi,\mu}(s)|<\frac{1}{2}\cdot\mathrm{gap}_{\min},$$

we can conclude that (π,μ) is a Nash Equilibrium, which means its duality regret is 0.

Proof of Proposition 1. We are going to prove $(\pi_h, \mu_h) = (\pi_h^*, \mu_h^*)\ \forall h \in [2H]$ by using the method of induction. When $h = 2H$, notice that for $\forall s \in \mathcal{S}$:

$$V_{2H}^{\pi,\mu}(s) = r(s, \mu_{2H}(s)) = Q_{2H}^*(s, \mu_{2H}(s)).$$

Since

$$\begin{aligned}\frac{1}{2} \cdot \text{gap}_{\min} &> |V_{2H}^*(s) - V_{2H}^{\pi,\mu}(s)| \\ &= |V_{2H}^*(s) - Q_{2H}^*(s, \mu_{2H}(s))|,\end{aligned}$$

therefore we have $\mu_{2H}^*(s) = \mu_{2H}(s)$ according to the definition of $\text{gap}_{\min}^+$. It means that (π, μ) and (π^*, μ^*) make the same decisions at the $2H$'s step. On the other hand, assume that $(\pi_t, \mu_t) = (\pi_t^*, \mu_t^*)$ holds for $t = h+1, h+2, \ldots, 2H$, then for $t = h$: if h is an odd number, then it's max-player's turn to take action, and it holds that for $\forall s \in \mathcal{S}$:

$$\begin{aligned}V_h^{\pi,\mu}(s) &= r(s, \pi(s)) + \mathbb{E}_{s'|s,\pi(s)} V_{h+1}^{\pi,\mu}(s') \\ &= r(s, \pi(s)) + \mathbb{E}_{s'|s,\pi(s)} V_{h+1}^*(s') = Q_h^*(s, \pi_h(s)).\end{aligned}$$

According to the following inequality:

$$\frac{1}{2} \cdot \text{gap}_{\min} > |V_h^*(s) - V_h^{\pi,\mu}(s)| = |V_h^*(s) - Q_h^*(s, \pi_h(s))|,$$

we can conclude that $\pi_h(s) = \pi_h^*(s)$ according to the definition of $\text{gap}_{\min}^+$. Therefore, (π, μ) and (π^*, μ^*) also make the same decisions at the h's step, which finishes our induction. When h is an even number, we can finish our induction in the same way, and that comes to the conclusion of Proposition 1 since $(\pi, \mu) = (\pi^*, \mu^*)$ and then the policy pair (π, μ) is a Nash Equilibrium. It leads to fact that:

$$V_1^{\dagger,\mu} = V_1^{\pi,\mu} = V_1^{\pi,\dagger},$$

which means policy pair (π, μ) has zero duality regret.

□

The conclusion shows that when vanilla regret is sufficiently small for all $h \in [2H], s \in \mathcal{S}$, then the policy pair is a Nash Equilibrium. It bridges the gap between the vanilla regret and the duality regret, and makes it more reasonable and convincing for us to work on the upper bound for the expected sum of vanilla regrets.

5 REGRET BOUND OF DISCOUNTED 2-TBSG

In this section, we study the discounted 2-TBSG with infinite horizon. In order to obtain a policy tuple sequence (pi^k, μ^k) to approximate a Nash equilibrium, we propose the following Algorithm 2, which is similar to Algorithm 1. We also use the UCB technique

Algorithm 2 Optimistic Nash Q-learning on Discounted 2-TBSG with Infinite Horizon

Initialize: Let $\overline{Q}(s,a), \hat{Q}(s,a) \leftarrow 1/(1-\gamma)$ and $\underline{Q}(s,a), \check{Q}(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Also $N(s,a) \leftarrow 0$.

Define: $\iota(k) = \log(SAT(k+1)(k+2)), \alpha_k = \frac{H+1}{H+k}$ where $H = \frac{\log(2/(1-\gamma)\text{gap}_{\min}^+)}{\log(1/\gamma)}$.

1: Observe the initial state s_1.
2: **for** episode $t \in [T]$ **do**
3: Take action $a_t \leftarrow \arg\max_{a' \in \mathcal{A}} \overline{Q}(s_t, a')$ if t is an odd number, (i.e. $I(a_t) = 1$), else take action $a_t \leftarrow \arg\min_{a' \in \mathcal{A}} \underline{Q}(s_t, a')$. After that, observe the reward $r(s_t, a_t)$ as well as the next state s_{t+1}.
4: $k = N(s_t, a_t) \leftarrow N(s_t, a_t) + 1$.
5: $b_k \leftarrow \frac{c_2}{1-\gamma}\sqrt{H\iota(k)/k}$, Here c_2 is a constant that can be set to $4\sqrt{2}$.
6: $\overline{Q}(s_t, a_t) \leftarrow (1-\alpha_k) \cdot \overline{Q}(s_t, a_t) + \alpha_k \cdot [r(s_t, a_t) + \gamma \hat{V}(s_{t+1}) + b_k]$.
7: $\underline{Q}(s_t, a_t) \leftarrow (1-\alpha_k) \cdot \underline{Q}(s_t, a_t) + \alpha_k \cdot [r(s_t, a_t) + \gamma \check{V}(s_{t+1}) - b_k]$.
8: $\hat{Q}(s_t, a_t) \leftarrow \min(\hat{Q}(s_t, a_t), \overline{Q}(s_t, a_t))$, $\check{Q}(s_t, a_t) \leftarrow \max(\check{Q}(s_t, a_t), \underline{Q}(s_t, a_t))$.
9: $\hat{V}(s_t) \leftarrow \hat{Q}(s_t, a_t), \check{V}(s_t) \leftarrow \check{Q}(s_t, a_t)$.
10: **end for**

to establish a upper estimation $\hat{Q}(s,a)$ and a lower estimation $\check{Q}(s,a)$ of the optimal Q-function $Q^*(s,a)$.

Since the V-function denotes the expected discounted sum of rewards given the initial state s and the Q function denotes the expected discounted sum of rewards given the initial state s and the initial action a, so they can be described as:

$$V^\pi(s) := \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^{i-1} \cdot r(s_i, a_i) \ : \ s_1 = s\right],$$

$$Q^\pi(s,a) := \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^{i-1} \cdot r(s_i, s_i) \ : \ s_1 = s, a_1 = a\right].$$

where γ is the discounted ratio. In the following statements, $\overline{Q}_t, \overline{V}_t$ stand for the $\overline{Q}, \overline{V}$ functions in the t-th iteration, and so does the other subscripts.

5.1 Sub-optimality Gap and the Splitting of Total Regret

Similar to the episodic setting, given $(s,a) \in \mathcal{S} \times \mathcal{A}$, define $\text{gap}(s,a)$ as:

$$\text{gap}(s,a) := V^*(s) - Q^*(s,a).$$

Notice that, when $a \in \mathcal{A}_1$, which means the action a is taken by max-player P_1, then $V^*(s) \geqslant Q^*(s,a) \Rightarrow \text{gap}(s,a) \geqslant 0$. In contrast, when $a \in \mathcal{A}_2$, which means

the action a is taken by min-player P_2, then $V^*(s) \leqslant Q^*(s,a) \Rightarrow \text{gap}(s,a) \leqslant 0$. Here, we can introduce a notation $\text{gap}^+(s,a) := |\text{gap}(s,a)|$ which stands for the absolute value of $\text{gap}(s,a)$. Also, we denote $\text{gap}^+_{\min}$ as the minimum non-zero absolute gap:

$$\text{gap}^+_{\min} := \min_{s,a}\{\text{gap}^+(s,a) \ : \ \text{gap}^+(s,a) \neq 0\} > 0.$$

In the main theorem proposed in the following section, we will estimate an upper bound of the expected total regret for the first T steps

$$\text{Regret}(T) := \sum_{t=1}^{T} \left|(V^* - V^{\pi_t,\mu_t})(s_t)\right|.$$

5.2 Main Theorem

In this section, we propose our main theorem in the infinite-horizon setting. Unlike dual gap regret (Xie et al., 2020), the total regret above proposed by Liu and Su (2020) follows the sample complexity definition in Sm (2003) and directly compares the actual value function and the value function from the first t iterations. Similar to the episodic setting, we can obtain a gap-dependent logarithmic upper bound for the expected total regret.

Theorem 2 (Main Theorem 2: Logarithmic Regret Bound of Q-learning for Infinite-horizon Discounted 2-TBSG)**.** *After using Algorithm 2, the expected total regret for infinite-horizon two-player turn-based stochastic game can be upper bounded by:*

$$\mathbb{E}[\text{Regret}(T)] \leqslant \mathcal{O}\left(\frac{SA}{\text{gap}^+_{\min}(1-\gamma)^5\log(1/\gamma)} \cdot \log\frac{SAT}{\text{gap}^+_{\min}(1-\gamma)}\right).$$

We put the whole proof of Theorem 2 into the appendix. Its proof has the same structure as that of Theorem 1. We need to first split the regret into expected sum of gaps, then use concentration properties and peeling technique to finish the whole proof.

6 EPISODIC 2-TBSG WITH LINEAR FUNCTION APPROXIMATION

In this section, we analyze the gap dependent total regret bound of episodic 2-TBSG under the linear function expression assumption. Here, we have two different settings, centralized version and independent version. As follows, we are going to introduce the centralized version and put the independent version into the appendix. Besides, gap dependent regret bound will be provided in both settings.

6.1 Centralized Setting and Algorithm

Assumption 1 (Assumption 4.1 from He et al. (2020))**.** *A Markov Game* $\text{MG}(2H, \mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ *is defined as linear when the probability transition kernels and the reward functions are linear with respect to a given feature map* $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *where* d *is the feature dimension. Specifically, for each* $h \in [2H]$*, there exists an unknown vector* $\mu_h \in \mathbb{R}^d$ *and unknown measures* $\theta_h = (\theta_h^{(1)}, \theta_h^{(2)}, \ldots, \theta_h^{(d)})$ *whose degree of freedom is* $|S| \times d$*, such that for* $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$*:*

$$\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \theta_h(s') \rangle \text{ and } r(s,a) = \langle \phi(s,a), \mu_h \rangle.$$

For simplicity, we assume that $\|\phi(s,a)\|_2 \leqslant 1$, $\|\mu_h\|_2 \leqslant \sqrt{d}$ *and* $\|\theta_h(\mathcal{S})\| \leqslant \sqrt{d}$.

In the centralized setting, a central controller controls both players, and this central controller's goal is to learn a Nash Equilibrium. In our algorithm, both the max-player and the min-player update their policies (π^k, μ^k) according to the history information. Under Assumption 1, we know that for any policy π, the action-value function $Q_h^\pi(s,a)$ is a linear function $\langle \phi(s,a), \theta_h^\pi \rangle$ with respect to the feature $\phi(s,a)$. In order to estimate the optimal action-value function $Q_h^*(s,a) := \langle \phi(s,a), \theta_h^* \rangle$, we only have to estimate the parameters θ_h^*. Since Assumption 1 only gives a condition on the linear structure of the stochastic game, it is still a two-player turn-based general sum stochastic game with finite horizon. Therefore, we have exactly the same definition on the cumulative regret (or total regret) as Section 4:

$$\text{Regret}(K) = \sum_{k=1}^{K} \left|\left(V_1^* - V_1^{\pi^k,\mu^k}\right)(s_1^k)\right|.$$

In the following Least Square Value Iteration on 2-TBSG (LSVI-2TBSG) algorithm, we introduce two new variables $\overline{w}_h^k, \underline{w}_h^k$, which are the upper and lower estimations of θ_h^* in the k-th episode. They are computed by solving the following regularized least-square problems:

$$\begin{aligned} \overline{w}_h^k &\leftarrow \arg\min_{w \in \mathbb{R}^d} \lambda\|w\|^2 + F_1(w), \\ \underline{w}_h^k &\leftarrow \arg\min_{w \in \mathbb{R}^d} \lambda\|w\|^2 + F_2(w), \end{aligned} \tag{7}$$

where

$$F_1(w) = \sum_{i=1}^{k-1}\left[\phi(s_h^i, a_h^i)^\top w - r_h(s_h^i, a_h^i) - \overline{V}_{h+1}^k(s_{h+1}^i)\right]^2,$$

$$F_2(w) = \sum_{i=1}^{k-1}\left[\phi(s_h^i, a_h^i)^\top w - r_h(s_h^i, a_h^i) - \underline{V}_{h+1}^k(s_{h+1}^i)\right]^2.$$

Actually, these two least-square problems can be easily solved as:

$$\overline{w}_h^k = (\Lambda_h^k)^{-1} \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \left[r_h(s_h^i, a_h^i) + \overline{V}_{h+1}^k(s_{h+1}^i) \right],$$

$$\underline{w}_h^k = (\Lambda_h^k)^{-1} \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \left[r_h(s_h^i, a_h^i) + \underline{V}_{h+1}^k(s_{h+1}^i) \right],$$

where $\Lambda_h^k = \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$. Then, we update the estimated Q-values by:

$$\overline{Q}_h^k(s,a) = \min(2H, \phi(s,a)^\top \overline{w}_h^k + \beta \cdot T(s,a)),$$
$$\underline{Q}_h^k(s,a) = \max(0, \phi(s,a)^\top \underline{w}_h^k - \beta \cdot T(s,a)),$$

where $T(s,a) = \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)}$ can be regarded as a UCB term. For the pseudo-code of LSVI-2TBSG algorithm, we leave it to the appendix.

6.2 Main Theorem for the Centralized Setting

In this section, we propose the main theorem under the linear function expression assumption in the centralized setting. When using LSVI-2TBSG algorithm, the expected total regret can be upper bounded by the following theorem:

Theorem 3 (Main Theorem 3: Logarithmic Regret Bound of LSVI-2TBSG (Centralized))**.** *Under Assumption 1, after using LSVI-2TBSG algorithm (centralized)*

$$\mathbb{E}[\text{Regret}(K)] \leqslant 1 + \frac{Cd^3H^5 \log(16dK^2(K+1)H^3)}{\text{gap}_{\min}^+} \iota,$$

where $\iota = \log\left(\frac{Cd^3H^4\log(4dKH)}{(\text{gap}_{\min}^+)^2}\right)$ *is a logarithmic term.*

By using similar techniques, we can prove the theorem and propose the first gap-dependent logarithmic regret bound under the linear MDP assumption. We leave the technical proof of this theorem to the appendix.

7 CONCLUSION

In this paper, we gave the first set of gap-dependent logarithmic regret bounds for two-player turn-based stochastic Markov games in both tabular and cases, and in both centralized setting and independent settings.

The current bound is tight on T but not tight on the horizon H, so one of our possible future directions is to further improve our bounds in terms of the dependency on horizon. Another fruitful future direction is to extend our analysis to more general settings (Wang et al., 2020; Jiang et al., 2017; Du et al., 2021; Jin et al., 2021), including the multi-player setting and the mean-field game setting.

8 Acknowledgements

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P.Morgan, and Two Sigma for their supports. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning). Simon Shaolei Du gratefully acknowledges the funding from NSF Award's IIS-2110170 and DMS-2134106.

References

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

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

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML'17 Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 263–272, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 551–560, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 5713–5723, 2017.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

Simon S. Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, volume 32, pages 8060–8070, 2019.

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.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

Mohammad Ghavamzadeh, Hilbert J. Kappen, Mohammad G. Azar, and Rémi Munos. Speedy q-learning. In *Advances in Neural Information Processing Systems 24*, volume 24, pages 2411–2419, 2011.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. *arXiv preprint arXiv:2011.11566*, 2020.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

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, Lin Yang, Csaba Szepesvari, Mengdi Wang, and Alex Ayoub. Model-based reinforcement learning with value-targeted regression. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 463–474, 2020.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *32nd Conference on Neural Information Processing Systems, NeurIPS 2018*, volume 31, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2019.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021.

Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems 11*, volume 11, pages 996–1002, 1998.

Sven Koenig and Reid G. Simmons. Complexity analysis of real-time reinforcement learning. In *AAAI'93 Proceedings of the eleventh national conference on Artificial intelligence*, pages 99–105, 1993.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *ALT'12 Proceedings of the 23rd international conference on Algorithmic Learning Theory*, pages 320–334, 2012.

Shuang Liu and Hao Su. Regret bounds for discounted mdps. *arXiv: Learning*, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. In *32nd Conference on Neural Information Processing Systems (NIPS), DEC 02-08, 2018, Montreal, CANADA*, volume 31, pages 8874–8882, 2018.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, pages 2377–2386, 2016.

Todd A Schultz. Linear complementarity and discounted switching controller stochastic games. *Journal of optimization theory and applications*, 73(1): 89–99, 1992.

L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 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, 2018.

Max Simchowitz and Kevin G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, volume 32, pages 1153–1162, 2019.

Kakade Sm. On the sample complexity of reinforcement learning. 2003.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. 1988.

Ambuj Tewari and Peter L. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems 20*, volume 20, pages 1505–1512, 2007.

Martin J. Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv: Learning*, 2019.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33, 2020.

Yining Wang, Ruosong Wang, Simon S. Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Christopher J.C.H. Watkins and Peter Dayan. Technical note q-learning. In *Machine Learning*, volume 8, pages 279–292, 1992.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. *arXiv preprint arXiv:2010.01374*, 2020.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Thirty-Third Annual Conference On Learning Theory*, pages 3674–3682, 2020.

Haike Xu, Tengyu Ma, and Simon S Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

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

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 10746–10756, 2020.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020.

SUPPLEMENTARY MATERIAL

A Proof of Lemma 4

According to the update rule of Algorithm 1, we can get the following equations: let $t = N_h^k(s,a)$ be the total times when state-action tuple (s,a) appears at the h-th step in the first $k-1$ episodes. Suppose tuple (s,a) previously appeared at episodes $k^1, k^2, \cdots, k^t < k$ at the h-th step. Denote $\tau_h(s,a,i) := k^i$. Then, it holds that:

$$\begin{aligned}
\overline{Q}_h^k(s,a) &= \alpha_t^0 \cdot 2H + \sum_{i=1}^{t} \alpha_t^i \left[r_h(s,a) + \overline{V}_{h+1}^{k^i}(s_{h+1}^{k^i}) + \beta_i \right], \\
\underline{Q}_h^k(s,a) &= \sum_{i=1}^{t} \alpha_t^i \left[r_h(s,a) + \underline{V}_{h+1}^{k^i}(s_{h+1}^{k^i}) - \beta_i \right].
\end{aligned} \tag{8}$$

Then, we can upper bound the weighted sum of upper-lower gaps.

$$\begin{aligned}
&\sum_{k=1}^{K} w_{k,h} \left(\overline{Q}_h^k - \underline{Q}_h^k \right)(s_h^k, a_h^k) \\
\leqslant &\sum_{k=1}^{K} w_{k,h} \left(\alpha_{n_h^k}^0 \cdot 2H + \sum_{i=1}^{n_h^k} \alpha_{n_h^k}^i \left(\overline{V}_{h+1}^{\tau_h(s,a,i)} - \underline{V}_{h+1}^{\tau_h(s,a,i)} \right)(s_{h+1}^{\tau(s,a,i)}) + 2\beta_{n_h^k} \right) \\
= &\sum_{k \leqslant K, n_h^k = 0} w_{k,h} \cdot 2H + \sum_{k=1}^{K} 2 w_{k,h} \beta_{n_h^k} + \sum_{k=1}^{K} \left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k \right)(s_{h+1}^k) \left(\sum_{i=n_h^k+1}^{N_h^k(s_h^k, a_h^k)} \alpha_i^{n_h^k} w_{\tau_h(s_h^k, a_h^k, i), h} \right) \\
\leqslant &\, 2SAHw + \sum_{k=1}^{K} 2 w_{k,h} \beta_{n_h^k} + \sum_{k=1}^{K} w_{k,h+1} \left(\overline{Q}_{h+1}^k - \underline{Q}_{h+1}^k \right)(s_{h+1}^k, a_{h+1}^k),
\end{aligned} \tag{9}$$

where

$$w_{k,h+1} = \sum_{i=n_h^k+1}^{N_h^k(s_h^k, a_h^k)} \alpha_i^{n_h^k} w_{\tau_h(s_h^k, a_h^k, i), h}.$$

We can prove that:

$$\sum_{k=1}^{K} w_{k,h} \beta_{n_h^k} \leqslant 10c\sqrt{SACw(2H)^3\iota}. \tag{10}$$

Also, we can prove that: $\{w_{k,h+1}\}_{k \in [K]}$ is a $\left(C, \left(1 + \frac{1}{2H}\right) w\right)$-sequence. Therefore, after reversing this argument for $h+1, h+2, \ldots, 2H$, we obtain the following inequality:

$$\begin{aligned}
\sum_{k=1}^{K} w_{k,h} \left(\overline{Q}_h^k - \underline{Q}_h^k \right)(s_h^k, a_h^k) &\leqslant \sum_{h'=0}^{2H-h} \left(2SAH \cdot \left(1 + \frac{1}{2H}\right)^{h'} w + 10c\sqrt{SAC \left(1 + \frac{1}{2H}\right)^{h'} w(2H)^3\iota} \right) \\
&\leqslant 2H \left(2SAHew + 10c\sqrt{SACew(2H)^3\iota} \right) \\
&= 4ewSAH^2 + 60c\sqrt{SACewH^5\iota}.
\end{aligned}$$

B Proof of Theorem 2

In this section, we are going to provide the whole proof of Theorem 2.

B.1 First step: Splitting the Regret into Expected Sum of Gaps

The splitting of the regret is just the same as that in the finite horizon setting:

$$\mathbb{E}[\text{Regret}(T)] = \mathbb{E}\left[\sum_{t=1}^{T}\left|\sum_{h=0}^{\infty}\gamma^h \text{gap}(s_{t+h}, a_{t+h})\right|\right] \leqslant \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h'=t}^{+\infty}\gamma^{h'-t}\text{gap}^+(s_{h'}, a_{h'})\right]. \tag{11}$$

Notice that when t is an odd number, it's the max-player's turn to take action, so $\text{gap}(s_t, a_t) \geqslant 0$ and then:

$$\begin{aligned}
&\text{gap}^+(s_t, a_t) = \text{gap}(s_t, a_t) = V^*(s_t) - Q^*(s_t, a_t) \\
&= Q^*(s_t, a^*) - Q^*(s_t, a_t) \leqslant \hat{Q}_t(s_t, a^*) - \breve{Q}_t(s_t, a_t) \\
&\leqslant \left(\hat{Q}_t - \breve{Q}_t\right)(s_t, a_t).
\end{aligned}$$

Similarly, when t is an even number, it's the min-player's turn to take action, so $\text{gap}(s_t, a_t) \leqslant 0$, and then:

$$\begin{aligned}
&\text{gap}^+(s_t, a_t) = -\text{gap}(s_t, a_t) = Q^*(s_t, a_t) - V^*(s_t) \\
&= Q^*(s_t, a_t) - Q^*(s_t, a^*) \leqslant \hat{Q}_t(s_t, a_t) - \breve{Q}_t(s_t, a^*) \\
&\leqslant \left(\hat{Q}_t - \breve{Q}_t\right)(s_t, a_t).
\end{aligned}$$

Therefore, we can conclude that $\text{gap}^+(s_t, a_t) \leqslant \left(\hat{Q}_t - \breve{Q}_t\right)(s_t, a_t)$. By the definition of $\text{gap}^+_{\min}$, we have:

$$\text{gap}^+(s_t, a_t) = \text{clip}[\text{gap}^+(s_t, a_t)|\text{gap}^+_{\min}] \leqslant \text{clip}\left[\left(\hat{Q}_t - \breve{Q}_t\right)(s_t, a_t)|\text{gap}^+_{\min}\right].$$

Combine it with Equation (11), we obtain that:

$$\mathbb{E}[\text{Regret}(T)] \leqslant \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h'=t}^{+\infty}\gamma^{h'-t}\cdot\text{clip}\left[\left(\hat{Q}_{h'} - \breve{Q}_{h'}\right)(s_{h'}, a_{h'})|\text{gap}^+_{\min}\right]\right]. \tag{12}$$

B.2 Second step: Concentration Properties

Extended from Dong et al. (2019), we can obtain the following lemma which shows that Algorithm 2 satisfies bounded learning error with high probability.

Lemma 5 (Concentration Property). *When applying Algorithm 2, event $\mathcal{E}_{\text{conc}}$ occurs with probability at least $1 - \frac{1}{T}$. Here, $\mathcal{E}_{\text{conc}}$ occurs if $\forall(s, a, t) \in \mathcal{S}\times\mathcal{A}\times\mathbb{N}$:*

$$0 \leqslant \left(\hat{Q}_t - Q^*\right)(s, a) \leqslant \left(\overline{Q}_t - Q^*\right)(s, a) \leqslant \frac{\alpha^0_{n^t}}{1-\gamma} + \sum_{i=1}^{n^t}\gamma\alpha^i_{n^t}\left(\hat{V}_{\tau(s,a,i)} - V^*\right)(s_{\tau(s,a,i)}) + \beta_{n^t},$$

$$0 \leqslant \left(Q^* - \breve{Q}_t\right)(s, a) \leqslant \left(Q^* - \underline{Q}_t\right)(s, a) \leqslant \frac{\alpha^0_{n^t}}{1-\gamma} + \sum_{i=1}^{n^t}\gamma\alpha^i_{n^t}\left(V^* - \breve{V}_{\tau(s,a,i)}\right)(s_{\tau(s,a,i)}) + \beta_{n^t}.$$

Here, $\iota(k) = \log(SAT(k+1)(k+2))$ and $\beta_k = \frac{c_3}{1-\gamma}\sqrt{\frac{H\iota(k)}{k}}$.

Then, under $\mathcal{E}_{\text{conc}}$, we know that:

$$0 \leqslant \left(\hat{Q}_t - \breve{Q}_t\right)(s, a) \leqslant \frac{2\alpha^0_{n^t}}{1-\gamma} + 2\beta_{n^t} + \sum_{i=1}^{n^t}\gamma\alpha^i_{n^t}\left(\hat{V}_{\tau(s,a,i)} - \breve{V}_{\tau(s,a,i)}\right)(s_{\tau(s,a,i)}). \tag{13}$$

B.3 Third step: Peeling

Similar to Lemma 4, we upper bound the weighted sum of upper-lower gaps and then bound the counting number of gaps in different intervals, just like we did in the episodic setting.

Lemma 6 (Peeling Argument)**.** *Under the event $\mathcal{E}_{\text{conc}}$, the following holds for any weighted sequence $\{\omega_t\}$ which satisfies: $0 \leqslant \omega_t \leqslant \omega, \sum_{t=1}^{+\infty} \omega_t \leqslant C$.*

$$\sum_{t=1}^{+\infty} \omega_t \left(\hat{Q}_t - \check{Q}_t\right)(s_t, a_t) \leqslant \frac{\gamma^H C}{1-\gamma} + \mathcal{O}\left(\frac{\sqrt{\omega SAHC\iota(C)} + \omega SA}{(1-\gamma)^2}\right).$$

After that, we classify the positive gaps into different intervals. Since all the positive gaps $\text{gap}^+(s,a) \in [\text{gap}^+_{\min}, 1/(1-\gamma))$ and this interval can be separated into N intervals

$$\Lambda_n = [2^{n-1}\text{gap}^+_{\min}, 2^n \text{gap}^+_{\min})$$

where $n = 1, 2, \ldots, N$. Here, $N = \left\lceil \log_2\left(\frac{1}{\text{gap}^+_{\min}(1-\gamma)}\right)\right\rceil$. Under the event $\mathcal{E}_{\text{conc}}$, for $n \in [N]$, we define:

$$C^{(n)} := \left|\left\{t \in \mathbb{N}_+ \ : \left(\hat{Q}_t - \check{Q}_t\right)(s_t, a_t) \in \Lambda_n\right\}\right|.$$

By using the sequence

$$\omega_t^{(n)} := \mathbb{I}\left[\left(\hat{Q}_t - \check{Q}_t\right)(s_t, a_t) \in \Lambda_n\right],$$

we can upper bound the $C^{(n)}$ by using Lemma 6.

Lemma 7. *We can upper bound the $C^{(n)}$ by:*

$$\mathcal{O}\left(\frac{SA}{4^n(\text{gap}^+_{\min})^2(1-\gamma)^4\log(1/\gamma)} \log\left(\frac{SAT}{(1-\gamma)\text{gap}^+_{\min}}\right)\right).$$

Finally, we come to our main theorem. According to Equation (12), we know that: if the trajectory satisfies the $\mathcal{E}_{\text{conc}}$ condition, then:

$$\begin{aligned}
\text{Regret}(T) &\leqslant \frac{1}{1-\gamma}\sum_{t=1}^{+\infty} \text{clip}\left[\left(\hat{Q}_t - \check{Q}_t\right)(s_t, a_t) | \text{gap}^+_{\min}\right] \leqslant \frac{1}{1-\gamma}\sum_{n=1}^{N} 2^n \text{gap}^+_{\min} C^{(n)} \\
&\leqslant \sum_{n=1}^{N} \mathcal{O}\left(\frac{SA}{2^n \text{gap}^+_{\min}(1-\gamma)^5\log(1/\gamma)} \cdot \iota\right) = \mathcal{O}\left(\frac{SA}{\text{gap}^+_{\min}(1-\gamma)^5\log(1/\gamma)} \cdot \iota\right). \qquad (14)
\end{aligned}$$

Here, $\iota = \log\left(\frac{SAT}{\text{gap}^+_{\min}(1-\gamma)}\right)$ is the logarithmic term on T. For the other trajectories outside $\mathcal{E}_{\text{conc}}$, we have a trivial upper bound:

$$\text{Regret}(T) \leqslant \sum_{t=1}^{T}\sum_{h'=t}^{+\infty} \gamma^{h'-t} \cdot \left(\hat{Q}_{h'} - \check{Q}_{h'}\right)(s_{h'}, a_{h'}) \leqslant \sum_{t=1}^{T}\sum_{h'=t}^{+\infty} \frac{\gamma^{h'-t}}{1-\gamma} \leqslant \frac{T}{(1-\gamma)^2}. \qquad (15)$$

Now we combine Equation (14) and Equation (15), we obtain that:

$$\begin{aligned}
\mathbb{E}[\text{Regret}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h'=t}^{+\infty} \gamma^{h'-t}\text{gap}(s_{h'}, a_{h'})\right] \\
&\leqslant \mathbb{P}(\overline{\mathcal{E}_{\text{conc}}}) \cdot \frac{T}{(1-\gamma)^2} + \mathbb{P}(\mathcal{E}_{\text{conc}}) \cdot \mathcal{O}\left(\frac{SA}{\text{gap}^+_{\min}(1-\gamma)^5\log(1/\gamma)} \cdot \iota\right) \\
&\leqslant \mathcal{O}\left(\frac{SA}{\text{gap}^+_{\min}(1-\gamma)^5\log(1/\gamma)} \cdot \iota\right). \qquad (16)
\end{aligned}$$

which comes from $\mathbb{P}(\overline{\mathcal{E}_{\text{conc}}}) \leqslant 1/T$. Theorem 2 is proved and it provides us an upper bound which is logarithmically dependent on T.

B.4 Proof of Lemma 6

By using the conclusion of Lemma 5, we know that:

$$\sum_{t=1}^{+\infty}\omega_t\left(\hat{Q}_t-\check{Q}_t\right)(s_t,a_t)\leqslant\sum_{t=1}^{+\infty}\frac{2\omega_t\alpha_{n^t}^0}{1-\gamma}+\sum_{t=1}^{+\infty}2\omega_t\beta_{n^t}+\gamma\sum_{t=1}^{+\infty}\sum_{i=1}^{n^t}\omega_t\alpha_{n^t}^i\left(\hat{V}_{\tau(s,a,i)}-\check{V}_{\tau(s,a,i)}\right)(s_{\tau(s,a,i)}).$$

Now, we analyze the three terms above one by one.

$$\sum_{t=1}^{+\infty}\frac{2\omega_t\alpha_{n^t}^0}{1-\gamma}\leqslant\sum_{t=1}^{+\infty}\mathbb{I}[n^t=0]\frac{2\omega}{1-\gamma}=\frac{2SA\omega}{1-\gamma},\tag{17}$$

$$\begin{aligned}&\sum_{t=1}^{+\infty}\omega_t\beta_{n^t}=\sum_{s,a}\sum_{i=1}^{N(s,a)}\omega_{\tau(s,a,i)}\beta_i=\frac{c_3\sqrt{H}}{1-\gamma}\sum_{s,a}\sum_{i=1}^{N(s,a)}\omega_{\tau(s,a,i)}\sqrt{\frac{\iota(i)}{i}}\\&\leqslant\frac{c_3\sqrt{H}}{1-\gamma}\sum_{s,a}\sum_{i=1}^{C_{s,a}/\omega}\omega\sqrt{\frac{\iota(C)}{i}}\leqslant\frac{2c_3\sqrt{H}}{1-\gamma}\sum_{s,a}\sqrt{C_{s,a}\omega\iota(C)}\leqslant\frac{2c_3}{1-\gamma}\sqrt{SAHC\omega\iota(C)}.\end{aligned}\tag{18}$$

Here, $C_{s,a}=\sum_{i=1}^{N(s,a)}\omega_{\tau(s,a,i)}$ is the partial sum of the sequence $\{\omega_t\}$, and therefore:

$$\sum_{s,a}C_{s,a}\leqslant C.$$

Equation 17 and Equation 18 stand for the first and second terms. Now we analyze the third term.

$$\begin{aligned}&\gamma\sum_{t=1}^{+\infty}\sum_{i=1}^{n^t}\omega_t\alpha_{n^t}^i\left(\hat{V}_{\tau(s,a,i)}-\check{V}_{\tau(s,a,i)}\right)(s_{\tau(s,a,i)})\\=&\gamma\sum_{t=1}^{+\infty}\left(\hat{V}_t-\check{V}_t\right)(s_{t+1})\sum_{i\geqslant n^t+1}\omega_{\tau(s_t,a_t,i)}\alpha_i^{n^t}\\=&\gamma\sum_{t=2}^{+\infty}\omega_t'\left(\hat{V}_t-\check{V}_t\right)(s_t)+\gamma\sum_{t=1}^{+\infty}\omega_{t+1}'\left(\hat{V}_t-\hat{V}_{t+1}\right)(s_{t+1})+\gamma\sum_{t=1}^{+\infty}\omega_{t+1}'\left(\check{V}_{t+1}-\check{V}_t\right)(s_{t+1}).\end{aligned}\tag{19}$$

Here:

$$\omega_{t+1}':=\sum_{i\geqslant n^t+1}\omega_{\tau(s_t,a_t,i)}\alpha_i^{n^t}$$

can be easily verified as a $(C,(1+1/H)\omega)$-sequence. By the update rule of Algorithm 2, $\hat{Q}_t(s,a)$ is decreasing and $\check{Q}_t(s,a)$ is increasing by t for $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$. Therefore, $\hat{(V)}_t(s)$ is decreasing and $\check{V}_t(s)$ is increasing by t for $\forall s\in\mathcal{S}$. Then:

$$\gamma\sum_{t=1}^{+\infty}\omega_{t+1}'\left(\hat{V}_t-\hat{V}_{t+1}\right)(s_{t+1})\leqslant\gamma(1+1/H)\omega\sum_s\sum_{t=1}^{+\infty}\left(\hat{V}_t-\hat{V}_{t+1}\right)(s)\leqslant\frac{\gamma(1+1/H)\omega S}{1-\gamma},$$

and similarly:

$$\gamma\sum_{t=1}^{+\infty}\omega_{t+1}'\left(\check{V}_{t+1}-\check{V}_t\right)(s_{t+1})\leqslant\gamma(1+1/H)\omega\sum_s\sum_{t=1}^{+\infty}\left(\check{V}_{t+1}-\check{V}_t\right)(s)\leqslant\frac{\gamma(1+1/H)\omega S}{1-\gamma}.$$

Therefore, Equation (19) leads to:

$$\gamma\sum_{t=1}^{+\infty}\sum_{i=1}^{n^t}\omega_t\alpha_{n^t}^i\left(\hat{V}_{\tau(s,a,i)}-\check{V}_{\tau(s,a,i)}\right)(s_{\tau(s,a,i)})\leqslant\gamma\sum_{t=2}^{+\infty}\omega_t'\left(\hat{Q}_t-\check{Q}_t\right)(s_t,a_t)+\frac{2\gamma(1+1/H)\omega S}{1-\gamma}.$$

After combining with Equation (17) and Equation (18), we obtain that:

$$\begin{aligned}
&\sum_{t=1}^{+\infty}\omega_t\left(\hat{Q}_t-\breve{Q}_t\right)(s_t,a_t)\\
\leqslant&\frac{2SA\omega}{1-\gamma}+\frac{2c_3}{1-\gamma}\sqrt{SAHC\omega\iota(C)}+\frac{2\gamma(1+1/H)\omega S}{1-\gamma}+\gamma\sum_{t=2}^{+\infty}\omega_t'\left(\hat{Q}_t-\breve{Q}_t\right)(s_t,a_t)\\
=&\mathcal{O}\left(\frac{SA\omega+\sqrt{SAHC\omega\iota(C)}}{1-\gamma}\right)+\gamma\sum_{t=2}^{+\infty}\omega_t'\left(\hat{Q}_t-\breve{Q}_t\right)(s_t,a_t).
\end{aligned}$$

We can repeat this unrolling argument for H times, and get a $(C,(1+1/H)^H\omega\leqslant e\omega)$-sequence $\{\omega_t^{(H)}\}_{t\geqslant H+1}$. Then, we get the following result.

$$\begin{aligned}
&\sum_{t=1}^{+\infty}\omega_t\left(\hat{Q}_t-\breve{Q}_t\right)(s_t,a_t)\\
=&\sum_{h=1}^{H}\gamma^h\mathcal{O}\left(\frac{SA\omega+\sqrt{SAHC\omega\iota(C)}}{1-\gamma}\right)+\gamma^H\sum_{t=H+1}^{+\infty}\omega_t^{(H)}\left(\hat{Q}_t-\breve{Q}_t\right)(s_t,a_t)\\
\leqslant&\frac{1}{1-\gamma}\cdot\mathcal{O}\left(\frac{SA\omega+\sqrt{SAHC\omega\iota(C)}}{1-\gamma}\right)+\frac{\gamma^H}{1-\gamma}\sum_{t=H+1}^{+\infty}\omega_t^{(H)}\\
=&\mathcal{O}\left(\frac{SA\omega+\sqrt{SAHC\omega\iota(C)}}{(1-\gamma)^2}\right)+\frac{\gamma^HC}{1-\gamma},
\end{aligned}$$

which comes to our conclusion.

B.5 Proof of Lemma 7

Since $C^{(n)}=\sum_{t=1}^{+\infty}\omega_t^{(n)}$ and $\{\omega_t^{(n)}\}$ is a $(C^{(n)},1)$-sequence. According to Lemma 6,

$$\begin{aligned}
(2^{n-1}\mathrm{gap}_{\min}^+)\cdot C^{(n)}\leqslant&\sum_{t=1}^{+\infty}\omega_t^{(n)}\left(\hat{Q}_t-\breve{Q}_t\right)(s_t,a_t)\leqslant\mathcal{O}\left(\frac{SA+\sqrt{SAHC^{(n)}\iota(C^{(n)})}}{(1-\gamma)^2}\right)+\frac{\gamma^HC^{(n)}}{1-\gamma}\\
=&\frac{\mathrm{gap}_{\min}^+}{2}C^{(n)}+\mathcal{O}\left(\frac{SA+\sqrt{SAHC^{(n)}\iota(C^{(n)})}}{(1-\gamma)^2}\right).
\end{aligned}$$

Here, we use the fact that $\gamma^H=\frac{\mathrm{gap}_{\min}^+(1-\gamma)}{2}$. Denote $C^{(n)}=SAC'$, then:

$$\begin{aligned}
&(2^{n-2}\mathrm{gap}_{\min}^+)\cdot C^{(n)}\leqslant(2^{n-1}-\frac{1}{2})\mathrm{gap}_{\min}^+C^{(n)}\leqslant O\left(\frac{SA+\sqrt{SAHC^{(n)}\iota(C^{(n)})}}{(1-\gamma)^2}\right)\\
\Rightarrow&(2^{n-2}\mathrm{gap}_{\min}^+)\cdot C'\leqslant O\left(\frac{1+\sqrt{HC'\iota(C^{(n)})}}{(1-\gamma)^2}\right)\leqslant O\left(\frac{1+\sqrt{HC'\log(SATC')}}{(1-\gamma)^2}\right).
\end{aligned}$$

After solving the inequality above, we obtain that:

$$C'\leqslant\mathcal{O}\left(\frac{\log\left(\frac{SAT}{\mathrm{gap}_{\min}^+(1-\gamma)}\right)}{4^n(\mathrm{gap}_{\min}^+)^2(1-\gamma)^4\log(1/\gamma)}\right).$$

Therefore,

$$C^{(n)}\leqslant\mathcal{O}\left(\frac{SA}{4^n(\mathrm{gap}_{\min}^+)^2(1-\gamma)^4\log(1/\gamma)}\cdot\log\left(\frac{SAT}{\mathrm{gap}_{\min}^+(1-\gamma)}\right)\right),$$

which comes to our conclusion.

C Independent Setting and Algorithm for 2-TBSG with Linear Function Expression

In the independent setting, we do not have a central controller who controls both players. We can only control the max-player and play against the min-player whose policies are arbitrary but potentially adversarial. Since we only control the max-player, our goal is not to learn a Nash Equilibrium, but to maximize the reward of the max-player. Because of the differences between the centralized setting and the independent setting, we are going to redefine the gaps and regret functions in this section.

Since we can not get access to the min-player's policies and the Markov model of the game a priori, we are interested in the exploitability of max-player:

$$\text{Explicit}(\pi^k, \mu^k) := V_1^{\dagger,\mu^k}(s_1^k) - V_1^{\pi^k,\mu^k}(s_1^k),$$

which measures how much better the max-player can perform. Then, the cumulative regret can be defined as the sum of exploitability in different episodes:

$$\text{Regret}_\mu(K) := \sum_{k=1}^{K} \text{Explicit}(\pi^k, \mu^k) = \sum_{k=1}^{K} \left(V_1^{\dagger,\mu^k}(s_1^k) - V_1^{\pi^k,\mu^k}(s_1^k) \right).$$

Also, we need to redefine the gap. Previously, the gap is defined as $\text{gap}_h(s,a) = |V_h^*(s) - Q_h^*(s,a)|$ and $\text{gap}_{\min}^+ := \min_{h,s,a}\{\text{gap}_h(s,a) > 0\}$. However, in the independent setting, we can not control the min-player so it is not suitable to only consider $(\pi, \mu) = (\pi^*, \mu^*)$ since the Nash Equilibrium point is our final target. In order to measure the gap caused by max-player, we define:

$$\text{gap}_h^\mu(s,a) = |V_h^{\dagger,\mu}(s) - Q_h^{\dagger,\mu}(s,a)|.$$

Here, we still need the absolute value since $V_h^{\dagger,\mu}(s) - Q_h^{\dagger,\mu}(s,a)$ is non-negative when h is an odd number (and it's the max-player's turn to take action) while non-positive when h is an even number. The minimal gap can be obtained after taking minimum over all the (h,s,a) tuples and all possible pure strategy μ:

$$\text{gap}_{\min}^+ := \min_{\mu,h,s,a}\{\text{gap}_h^\mu(s,a) > 0\}.$$

Notice that the total number of pure strategies is finite, so the minimal gap above is positive and well-defined. Similar as the LSVI-2TBSG algorithm, we introduce a new variable w_h^k, which is the upper estimation of θ_h^* in the k-th episode. Here, we do not need the lower estimation $\underline{w}_h^k$ since the min-player's policy is beyond our control. In each episode, the w_h^k is computed by solving the following regularized least-square problem:

$$w_h^k \leftarrow \arg\min_{w\in\mathbb{R}^d} \lambda\|w\|^2 + \sum_{i=1}^{k-1} \left[\phi(s_h^i, a_h^i)^\top w - r_h(s_h^i, a_h^i) - \overline{V}_{h+1}^k(s_{h+1}^i) \right]^2.$$

Actually, it can be solved as:

$$w_h^k = (\Lambda_h^k)^{-1} \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \left[r_h(s_h^i, a_h^i) + \overline{V}_{h+1}^k(s_{h+1}^i) \right],$$

where $\Lambda_h^k = \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$. Then, we update the estimated Q-values by:

$$\overline{Q}_h^k(s,a) = \min(2H, \phi(s,a)^\top \overline{w}_h^k + \beta \cdot T(s,a)),$$

where $T(s,a) = \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)}$ is the UCB term. We leave the pseudo-code of the algorithm to the appendix.

C.1 Main Theorem for the Independent Setting

In this section, we propose the main theorem under the linear function expression assumption in the independent setting. The expected cumulative regret can be upper bounded by a logarithmic term:

Theorem 4 (Main Theorem 4: Logarithmic Regret Bound of LSVI-2TBSG (Independent)). *Under Assumption 1, after using LSVI-2TBSG algorithm (independent)*

$$\mathbb{E}[\text{Regret}(K)] \leqslant 1 + \frac{Cd^3H^5\log(16dK^2(K+1)H^3)}{\text{gap}_{\min}^+}\iota,$$

where $\iota = \log\left(\frac{Cd^3H^4\log(4dKH)}{(\text{gap}_{\min}^+)^2}\right)$ *is a logarithmic term.*

We provide a logarithmic regret upper bound for 2-TBSG under the linear MDP assumption, in both centralized and independent settings. To the best of our knowledge, this is the very first gap dependent upper bound for 2-TBSG, which makes our results novel and complete.

D Proofs for 2-TBSG with Linear Function Expression

In this section, we will give a theoretical proof on the Theorem 3. First, we prove some common lemmas of both settings.

D.1 Concentration Properties

The concentration property is important in controlling the fluctuations through the iterations. First, we introduce the following three lemmas proposed by Jin et al. (2019).

Lemma 8 (Lemma B.3 of Jin et al. (2019)). *Under Assumption 1, there exists an absolute constant C that is independent of c_β, such that with probability at least $1-p$, the following event $\mathcal{E}_{\text{conc}}$ holds: For centralized setting,*

$$\forall(k,h)\in[K]\times[2H]: \quad \left\|\sum_{i=1}^{k-1}\phi_h^i[\overline{V}_{h+1}^k(s_{h+1}^i)-\mathbb{P}_h\overline{V}_{h+1}^k(s_h^i,a_h^i)]\right\|_{(\Lambda_h^k)^{-1}} \leqslant C\cdot dH\sqrt{\theta},$$
$$\left\|\sum_{i=1}^{k-1}\phi_h^i[\underline{V}_{h+1}^k(s_{h+1}^i)-\mathbb{P}_h\underline{V}_{h+1}^k(s_h^i,a_h^i)]\right\|_{(\Lambda_h^k)^{-1}} \leqslant C\cdot dH\sqrt{\theta},$$

For the independent setting,

$$\forall(k,h)\in[K]\times[2H]: \quad \left\|\sum_{i=1}^{k-1}\phi_h^i[\overline{V}_{h+1}^k(s_{h+1}^i)-\mathbb{P}_h\overline{V}_{h+1}^k(s_h^i,a_h^i)]\right\|_{(\Lambda_h^k)^{-1}} \leqslant C\cdot dH\sqrt{\theta},$$

where $\theta = \log[2(1+c_\beta)dT/p]$ *and* $\Lambda_h^k = \phi(s_h^k,a_h^k)$.

By using Lemma 8, we can upper bound the difference between the optimal Q-function values and the Q-function values proposed by Algorithm 3.

Lemma 9 (Lemma B.4 of Jin et al. (2019)). *For any fixed policy π, with the event $\mathcal{E}_{\text{conc}}$ holds, we can conclude that $\forall(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[2H]\times[K]$: For centralized setting:*

$$\langle\phi(s,a),\overline{w}_h^k\rangle - Q_h^\pi(s,a) = \mathbb{P}_h(\overline{V}_{h+1}^k - V_{h+1}^\pi)(s,a) + \overline{\Delta}_h^k(s,a)$$
$$\langle\phi(s,a),\underline{w}_h^k\rangle - Q_h^\pi(s,a) = \mathbb{P}_h(\underline{V}_{h+1}^k - V_{h+1}^\pi)(s,a) + \underline{\Delta}_h^k(s,a).$$

Here, $|\overline{\Delta}_h^k(s,a)|, |\underline{\Delta}_h^k(s,a)| \leqslant \beta\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}$. *If we make $\pi=(\pi^*,\mu^*)$, we have:*

$$\langle\phi(s,a),\overline{w}_h^k\rangle - Q_h^*(s,a) = \mathbb{P}_h(\overline{V}_{h+1}^k - V_{h+1}^*)(s,a) + \overline{\Delta}_h^k(s,a)$$
$$\langle\phi(s,a),\underline{w}_h^k\rangle - Q_h^*(s,a) = \mathbb{P}_h(\underline{V}_{h+1}^k - V_{h+1}^*)(s,a) + \underline{\Delta}_h^k(s,a),$$

and therefore,

$$\langle\phi(s,a),\overline{w}_h^k\rangle-\langle\phi(s,a),\underline{w}_h^k\rangle=\mathbb{P}_h(\overline{V}_{h+1}^k-\underline{V}_{h+1}^k)(s,a)+\overline{\Delta}_h^k(s,a)-\underline{\Delta}_h^k(s,a).$$

For the independent setting:

$$\langle\phi(s,a),w_h^k\rangle-Q_h^\pi(s,a)=\mathbb{P}_h(\overline{V}_{h+1}^k-V_{h+1}^\pi)(s,a)+\overline{\Delta}_h^k(s,a)$$

where $|\overline{\Delta}_h^k(s,a)|\leqslant\beta\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}$. *If we make* $\pi=(\mathrm{br}(\mu^k),\mu^k):=(\dagger,\mu^k)$ *and* $\pi=(\pi^*,\mu^*)$, *we have:*

$$\langle\phi(s,a),w_h^k\rangle-Q_h^{\dagger,\mu^k}(s,a)=\mathbb{P}_h(\overline{V}_{h+1}^k-V_{h+1}^{\dagger,\mu^k})(s,a)+\overline{\Delta}_h^k(s,a),$$
$$\langle\phi(s,a),w_h^k\rangle-Q_h^*(s,a)=\mathbb{P}_h(\overline{V}_{h+1}^k-V_{h+1}^*)(s,a)+\overline{\Delta}_h^k(s,a).$$

Finally, by using the method of induction, we obtain the following lemma, which shows that for centralized setting, $\overline{Q},\underline{Q}$ are the upper and lower bounds of Q^* respectively. For the independent setting, $\overline{Q}$ is the upper bound for every $Q^{\dagger,\mu^k}$.

Lemma 10 (Lemma B.5 of Jin et al. (2019)). *On the event* $\mathcal{E}_{\mathrm{conc}}$ *proposed in Lemma 8, we have: for the centralized setting,* $\forall(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[2H]\times[K]$*:*

$$\overline{Q}_h^k(s,a)\geqslant Q_h^*(s,a)\geqslant\underline{Q}_h^k(s,a).$$

For the independent setting, $\forall(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[2H]\times[K]$*:*

$$\overline{Q}_h^k(s,a)\geqslant Q_h^{\dagger,\mu^k}(s,a)\geqslant Q_h^*(s,a).$$

We also need the following technical lemma.

Lemma 11 (Lemma 6.5 of He et al. (2020)). *When* $\lambda=1$, *for any subset* $C=\{c_1,c_2,\ldots,c_k\}\subseteq[K]$ *and any* $h\in[2H]$, *we have:*

$$\sum_{i=1}^k(\phi_h^{c_i})^\top(\Lambda_h^{c_i})^{-1}\phi_h^{c_i}\leqslant 2d\log(1+k).$$

D.2 Classifying Positive Gaps into Intervals (Centralized version)

Consider the term $\overline{Q}_h^k(s_h^k,a_h^k)-\underline{Q}_h^k(s_h^k,a_h^k)$. Under the event $\mathcal{E}_{\mathrm{conc}}$, when h is odd:

$$\begin{aligned}\mathrm{gap}_h^+(s_h^k,a_h^k)&=V_h^*(s_h^k)-Q_h^*(s_h^k,a_h^k)\leqslant Q_h^*(s_h^k,a^*)-\underline{Q}_h^k(s_h^k,a_h^k)\\&\leqslant\overline{Q}_h^k(s_h^k,a^*)-\underline{Q}_h^k(s_h^k,a_h^k)\leqslant\overline{Q}_h^k(s_h^k,a_h^k)-\underline{Q}_h^k(s_h^k,a_h^k)\end{aligned}$$

and when h is even:

$$\begin{aligned}\mathrm{gap}_h^+(s_h^k,a_h^k)&=Q_h^*(s_h^k,a_h^k)-V_h^*(s_h^k)\leqslant Q_h^*(s_h^k,a_h^k)-Q_h^*(s_h^k,a^*)\\&\leqslant\overline{Q}_h^k(s_h^k,a_h^k)-\underline{Q}_h^k(s_h^k,a^*)\leqslant\overline{Q}_h^k(s_h^k,a_h^k)-\underline{Q}_h^k(s_h^k,a_h^k).\end{aligned}$$

Here, we've applied Lemma 10, and the way of choosing a_h^k. Since for $\forall(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[2H]\times[K]$, the $\mathrm{gap}_h^+(s_h^k,a_h^k)\leqslant\overline{Q}_h^k(s_h^k,a_h^k)-\underline{Q}_h^k(s_h^k,a_h^k)\leqslant 2H$. So we can classify all these gaps into different intervals. According to the definition of $\mathrm{gap}_{\min}^+$, if one gap $\mathrm{gap}_h^+(s_h^k,a_h^k)$ belongs to $[0,\mathrm{gap}_{\min}^+)$, then it must be 0. For the other gaps, we divide them into N different intervals $[\mathrm{gap}_{\min}^+,2\mathrm{gap}_{\min}^+),\ldots,[2^{N-1}\mathrm{gap}_{\min}^+,2^N\mathrm{gap}_{\min}^+)$. Here, $N=\lceil\log_2(2H/\mathrm{gap}_{\min}^+)\rceil$. Then, we obtain the following conclusion:

$$\mathbb{E}[\mathrm{Regret}(K)]\leqslant\sum_{k=1}^K\mathbb{E}\left[\sum_{h=1}^{2H}\mathrm{gap}_h^+(s_h^k,a_h^k)\right]\leqslant\sum_{h=1}^{2H}\mathbb{E}\left[\sum_{n=1}^N 2^n\mathrm{gap}_{\min}^+\cdot\mathcal{T}_h^{(n)}\right].\tag{20}$$

where $\mathcal{T}_k^{(n)}=\sum_{k=1}^K\mathbb{I}\left[\mathrm{gap}_h^+(s_h^k,a_h^k)\in[2^{n-1}\mathrm{gap}_{\min}^+,2^n\mathrm{gap}_{\min}^+)\right]$, which is the number of positive gaps at the h-th step that belongs to the interval $[2^{n-1}\mathrm{gap}_{\min}^+,2^n\mathrm{gap}_{\min}^+)$ during the first K episodes.

D.3 Classifying Positive Gaps into Intervals (Independent version)

For the independent setting, notice that:

$$\mathbb{E}[\text{Regret}_\mu(K)] = \sum_{k=1}^{K} \mathbb{E}\left[V_1^{\dagger,\mu^k}(s_1^k) - V_1^{\pi^k,\mu^k}(s_1^k)\right] = \sum_{k=1}^{K} \mathbb{E}\left[\sum_{\substack{1\leqslant h\leqslant 2H \\ h \text{ odd}}} \text{gap}_h^\mu(s_h^k, a_h^k) | \pi^k, \mu^k\right].$$

Consider the term $\overline{Q}_h^k(s_h^k, a_h^k) - Q_h^*(s_h^k, a_h^k)$. Under the event $\mathcal{E}_{\text{conc}}$, when h is odd:

$$\begin{aligned}\text{gap}_h^\mu(s_h^k, a_h^k) &= V_h^{\dagger,\mu^k}(s_h^k) - Q_h^{\dagger,\mu^k}(s_h^k, a_h^k) = Q_h^{\dagger,\mu^k}(s_h^k, \hat{a}) - Q_h^{\dagger,\mu^k}(s_h^k, a_h^k) \\ &\leqslant \overline{Q}_h^k(s_h^k, \hat{a}) - Q_h^*(s_h^k, a_h^k) \leqslant \overline{Q}_h^k(s_h^k, a_h^k) - Q_h^*(s_h^k, a_h^k).\end{aligned}$$

Here, we've applied Lemma 10, and the way of choosing a_h^k. Since in the regret decomposition above, we only need the gap with odd steps, we have:

$$\mathbb{E}[\text{Regret}(K)] \leqslant \sum_{k=1}^{K} \mathbb{E}\left[\sum_{h=1}^{2H} \left(\overline{Q}_h^k(s_h^k, a_h^k) - Q_h^*(s_h^k, a_h^k)\right)\right].$$

Since for $\forall (s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [2H] \times [K]$, the $\text{gap}_h^\mu(s_h^k, a_h^k) \leqslant \overline{Q}_h^k(s_h^k, a_h^k) - Q_h^*(s_h^k, a_h^k) \leqslant 2H$ when h is an odd number. So we can classify all these non-zero gaps into N different intervals $[\text{gap}_{\min}^+, 2\text{gap}_{\min}^+), \ldots, [2^{N-1}\text{gap}_{\min}^+, 2^N \text{gap}_{\min}^+)$. Here, $N = \lceil \log_2(2H/\text{gap}_{\min}^+) \rceil$. Then, we obtain the following conclusion:

$$\mathbb{E}[\text{Regret}(K)] \leqslant \sum_{k=1}^{K} \mathbb{E}\left[\sum_{\substack{1\leqslant h\leqslant 2H \\ h \text{ odd}}} \text{gap}_h^\mu(s_h^k, a_h^k)\right] \leqslant \sum_{h=1}^{2H} \mathbb{E}\left[\sum_{n=1}^{N} 2^n \text{gap}_{\min}^+ \cdot \mathcal{T}_h^{(n)}\right]. \tag{21}$$

where $\mathcal{T}_k^{(n)} = \sum_{k=1}^{K} \mathbb{I}\left[\text{gap}_h^\mu(s_h^k, a_h^k) \in [2^{n-1}\text{gap}_{\min}^+, 2^n \text{gap}_{\min}^+), h \text{ is odd}\right]$, which is the number of positive gaps at the h-th step that belongs to the interval $[2^{n-1}\text{gap}_{\min}^+, 2^n \text{gap}_{\min}^+)$ during the first K episodes. In the next two sections, we will upper bound the counting number $\mathcal{T}_h^{(n)}$, which is the final step of our proof. By using Equation (20, 21), we can upper bound the expected total regret in both centralized and independent settings.

D.4 Upper Bounding the Counting Number (Centralized version)

For a fixed $h \in [2H]$ and $n \leqslant \lceil \log_2(2H/\text{gap}_{\min}^+) \rceil$, we will upper bound the $\mathcal{T}_k^{(n)}$. Under $\mathcal{E}_{\text{conc}}$, denote:

$$\{k \in [K] \ : \ \text{gap}_h^+(s_h^k, a_h^k) \in [2^{n-1}\text{gap}_{\min}^+, 2^n \text{gap}_{\min}^+)\} = \{k_1, k_2, \ldots, k_t\} := \mathcal{D},$$

where $t = \mathcal{T}_h^{(n)}$, then consider the sum $\sum_{i=1}^{t} [\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k)]$. On one hand, this sum has a lower bound:

$$\sum_{k\in\mathcal{D}} [\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k)] \geqslant \sum_{k\in\mathcal{D}} \text{gap}_h^+(s_h^k, a_h^k) \geqslant 2^{n-1}\text{gap}_{\min}^+ \cdot \mathcal{T}_h^{(n)}. \tag{22}$$

On the other hand, we can also establish an upper bound. By using Lemma 9:

$$\begin{aligned}&\sum_{k\in\mathcal{D}} [\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k)] \leqslant \sum_{k\in\mathcal{D}} \left[2\beta\sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} + \phi(s_h^k, a_h^k)^\top (\overline{w}_h^k - \underline{w}_h^k)\right] \\ \leqslant\ & 2\beta \sum_{k\in\mathcal{D}} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} + \sum_{k\in\mathcal{D}} \left[\phi(s_h^k, a_h^k)^\top \overline{w}_h^k - \phi(s_h^k, a_h^k)^\top \underline{w}_h^k\right] \\ \leqslant\ & 2\beta \sum_{k\in\mathcal{D}} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} + \sum_{k\in\mathcal{D}} \left[\mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) + 2\beta\|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}\right]\end{aligned}$$

$$= 4\beta \sum_{k\in\mathcal{D}} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} + \sum_{k\in\mathcal{D}} [\overline{V}_{h+1}^k(s_{h+1}^k) - \underline{V}_{h+1}^k(s_{h+1}^k)] + \sum_{k\in\mathcal{D}} \varepsilon_h^k$$
$$= 4\beta \sum_{k\in\mathcal{D}} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} + \sum_{k\in\mathcal{D}} [\overline{Q}_{h+1}^k(s_{h+1}^k, a_{h+1}^k) - \underline{Q}_{h+1}^k(s_{h+1}^k, a_{h+1}^k)] + \sum_{k\in\mathcal{D}} \varepsilon_h^k. \tag{23}$$

Taking summation over $h' = h, h+1, \ldots, 2H$, we have:

$$\sum_{k\in\mathcal{D}} [\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k)] \leqslant 4\beta \sum_{h'=h}^{2H} \sum_{k\in\mathcal{D}} \|\phi(s_{h'}^k, a_{h'}^k)\|_{(\Lambda_{h'}^k)^{-1}} + \sum_{h'=h}^{2H} \sum_{k\in\mathcal{D}} \varepsilon_{h'}^k. \tag{24}$$

Here, $\varepsilon_h^k = \mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k) - (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_{h+1}^k)$, which forms a martingale difference sequence. For each $k \in [K]$, with probability at least $1-p$, it holds that:

$$\sum_{i=1}^{k} \sum_{h'=h}^{2H} \varepsilon_{h'}^i \leqslant \sqrt{8kH^2 \log(2/p)}.$$

After taking a union bound for all $k \in [K]$, we know that, with probability at least $1 - Kp$, it holds that:

$$\sum_{k\in\mathcal{D}} \sum_{h'=h}^{2H} \varepsilon_h^i \leqslant \sqrt{8\mathcal{T}_{h'}^{(n)} H^2 \log(2/p)}. \tag{25}$$

According to Lemma 11 and Cauchy Inequality, we have:

$$\begin{aligned} \sum_{k\in\mathcal{D}} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} &\leqslant \sqrt{\mathcal{T}_h^{(n)}} \cdot \sqrt{\sum_{k\in\mathcal{D}} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}^2} = \sqrt{\mathcal{T}_h^{(n)}} \cdot \sqrt{\sum_{k\in\mathcal{D}} (\phi_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k} \\ &\leqslant \sqrt{2d\mathcal{T}_h^{(n)} \cdot \log(1 + \mathcal{T}_h^{(n)})}. \end{aligned} \tag{26}$$

Combine the martingale concentration with $\mathcal{E}_{\text{conc}}$, by using Equation (24, 25, 26), we have: with probability at least $1 - (K+1)p$,

$$\sum_{k\in\mathcal{D}} [\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k)] \leqslant 4\beta \sqrt{2d\mathcal{T}_h^{(n)} \cdot \log(1 + \mathcal{T}_h^{(n)})} + \sqrt{8\mathcal{T}_{h'}^{(n)} H^2 \log(2/p)}. \tag{27}$$

Finally, by Equation (22) and Equation (27), we know that with probability at least $1 - (K+1)p$:

$$\begin{aligned} 2^{n-1}\text{gap}_{\min}^+ \cdot \mathcal{T}_h^{(n)} &\leqslant 2H \cdot 4\beta \sqrt{2d\mathcal{T}_h^{(n)} \cdot \log(1 + \mathcal{T}_h^{(n)})} + \sqrt{8\mathcal{T}_{h'}^{(n)} H^2 \log(2/p)} \\ &= 8c_\beta dH^2 \sqrt{\log(4dKH/p)} \cdot \sqrt{2d\mathcal{T}_h^{(n)} \cdot \log(1 + \mathcal{T}_h^{(n)})} + \sqrt{8\mathcal{T}_{h'}^{(n)} H^2 \log(2/p)}. \end{aligned}$$

Therefore, we conclude that there exists an absolute constant C such that:

$$\mathcal{T}_h^{(n)} \leqslant \frac{Cd^3H^4 \log(4dKH/p)}{4^n (\text{gap}_{\min}^+)^2} \cdot \log\left(\frac{Cd^3H^4 \log(4dKH/p)}{4^n (\text{gap}_{\min}^+)^2} \right). \tag{28}$$

D.5 Upper Bounding the Counting Number (Independent version)

For a fixed odd number $h \in [2H]$ and $n \leqslant \lceil \log_2(2H/\text{gap}_{\min}^+) \rceil$, we will upper bound the $\mathcal{T}_k^{(n)}$. Under $\mathcal{E}_{\text{conc}}$, again we denote:

$$\{k \in [K] \ : \ \text{gap}_h^+(s_h^k, a_h^k) \in [2^{n-1}\text{gap}_{\min}^+, 2^n \text{gap}_{\min}^+)\} = \{k_1, k_2, \ldots, k_t\} := \mathcal{D},$$

where $t = \mathcal{T}_h^{(n)}$, then consider the sum $\sum_{i=1}^t [\overline{Q}_h^k(s_h^k, a_h^k) - \underline{Q}_h^k(s_h^k, a_h^k)]$. On one hand, this sum has a lower bound:

$$\sum_{k\in\mathcal{D}} [\overline{Q}_h^k(s_h^k, a_h^k) - Q_h^*(s_h^k, a_h^k)] \geqslant \sum_{k\in\mathcal{D}} \text{gap}_h^+(s_h^k, a_h^k) \geqslant 2^{n-1}\text{gap}_{\min}^+ \cdot \mathcal{T}_h^{(n)}. \tag{29}$$

On the other hand, we can also establish an upper bound. By using Lemma 9:

$$\begin{aligned}
&\sum_{k\in\mathcal{D}}[\overline{Q}_h^k(s_h^k,a_h^k)-Q_h^*(s_h^k,a_h^k)] \leqslant \sum_{k\in\mathcal{D}}\left[\beta\sqrt{\phi(s_h^k,a_h^k)^\top(\Lambda_h^k)^{-1}\phi(s_h^k,a_h^k)}+\langle\phi(s_h^k,a_h^k),w_h^k\rangle-Q_h^*(s_h^k,a_h^k)\right]\\
&\leqslant 2\beta\sum_{k\in\mathcal{D}}\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}}+\sum_{k\in\mathcal{D}}\mathbb{P}_h(\overline{V}_{h+1}^k-V_{h+1}^*)(s_h^k,a_h^k)\\
&=2\beta\sum_{k\in\mathcal{D}}\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}}+\sum_{k\in\mathcal{D}}[\overline{V}_{h+1}^k(s_{h+1}^k)-V_{h+1}^*(s_{h+1}^k)]+\sum_{k\in\mathcal{D}}\varepsilon_h^k\\
&\leqslant 2\beta\sum_{k\in\mathcal{D}}\|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}}+\sum_{k\in\mathcal{D}}[\overline{Q}_{h+1}^k(s_{h+1}^k,a_{h+1}^k)-Q_{h+1}^*(s_{h+1}^k,a_{h+1}^k)]+\sum_{k\in\mathcal{D}}\varepsilon_h^k. \qquad (30)
\end{aligned}$$

The final step holds since

$$V_{h+1}^*(s_{h+1}^k)=\max_a Q_{h+1}^*(s_{h+1}^k,a)\geqslant Q_{h+1}^*(s_{h+1}^k,a_{h+1}^k).$$

After taking summation over $h'=h,h+1,\ldots,2H$, we have:

$$\sum_{k\in\mathcal{D}}[\overline{Q}_h^k(s_h^k,a_h^k)-Q_h^*(s_h^k,a_h^k)]\leqslant 2\beta\sum_{h'=h}^{2H}\sum_{k\in\mathcal{D}}\|\phi(s_{h'}^k,a_{h'}^k)\|_{(\Lambda_{h'}^k)^{-1}}+\sum_{h'=h}^{2H}\sum_{k\in\mathcal{D}}\varepsilon_{h'}^k. \qquad (31)$$

Here, $\varepsilon_h^k=\mathbb{P}_h(\overline{V}_{h+1}^k-V_{h+1}^*)(s_h^k,a_h^k)-(\overline{V}_{h+1}^k-V_{h+1}^*)(s_{h+1}^k)$, which forms a martingale difference sequence. For each $k\in[K]$, with probability at least $1-p$, it holds that:

$$\sum_{i=1}^{k}\sum_{h'=h}^{2H}\varepsilon_{h'}^i\leqslant\sqrt{8kH^2\log(2/p)}.$$

The following is exact the same as the centralized version. We take a union bound for all $k\in[K]$, we know that, with probability at least $1-Kp$, it holds that:

$$\sum_{k\in\mathcal{D}}\sum_{h'=h}^{2H}\varepsilon_h^i\leqslant\sqrt{8\mathcal{T}_{h'}^{(n)}H^2\log(2/p)}. \qquad (32)$$

Combine the martingale concentration with $\mathcal{E}_{\text{conc}}$, by using Equations (31), (32), (26), with probability at least $1-(K+1)p$, we have

$$\sum_{k\in\mathcal{D}}[\overline{Q}_h^k(s_h^k,a_h^k)-Q_h^*(s_h^k,a_h^k)]\leqslant 2\beta\sqrt{2d\mathcal{T}_h^{(n)}\cdot\log(1+\mathcal{T}_h^{(n)})}+\sqrt{8\mathcal{T}_{h'}^{(n)}H^2\log(2/p)}. \qquad (33)$$

Finally, by Equation (33), we know that with probability at least $1-(K+1)p$:

$$\begin{aligned}
2^{n-1}\text{gap}_{\min}^+\cdot\mathcal{T}_h^{(n)}&\leqslant 2H\cdot 2\beta\sqrt{2d\mathcal{T}_h^{(n)}\cdot\log(1+\mathcal{T}_h^{(n)})}+\sqrt{8\mathcal{T}_{h'}^{(n)}H^2\log(2/p)}\\
&=4c_\beta dH^2\sqrt{\log(4dKH/p)}\cdot\sqrt{2d\mathcal{T}_h^{(n)}\cdot\log(1+\mathcal{T}_h^{(n)})}+\sqrt{8\mathcal{T}_{h'}^{(n)}H^2\log(2/p)}.
\end{aligned}$$

Therefore, we conclude that there exists an absolute constant C such that:

$$\mathcal{T}_h^{(n)}\leqslant\frac{Cd^3H^4\log(4dKH/p)}{4^n(\text{gap}_{\min}^+)^2}\cdot\log\left(\frac{Cd^3H^4\log(4dKH/p)}{4^n(\text{gap}_{\min}^+)^2}\right). \qquad (34)$$

D.6 Final Step of the Proof

Here, we come to final step of our proof. For the centralized setting, we combine Equation (20) with the upper bound of counting numbers (Equation (28)), we have:

$$\begin{aligned}
\mathbb{E}[\text{Regret}(K)] &\leqslant \sum_{h=1}^{2H} \mathbb{E}\left[\sum_{n=1}^{N} 2^n \text{gap}_{\min}^{+} \cdot \mathcal{T}_h^{(n)}\right] \\
&\leqslant (K+1)p \cdot 4H^2K + \sum_{n=1}^{N} 2^n \text{gap}_{\min}^{+} \cdot \frac{Cd^3H^5\log(4dKH/p)}{4^n(\text{gap}_{\min}^{+})^2} \cdot \log\left(\frac{Cd^3H^4\log(4dKH/p)}{4^n(\text{gap}_{\min}^{+})^2}\right) \\
&\leqslant 4H^2K(K+1)p + \frac{Cd^3H^5\log(4dKH/p)}{\text{gap}_{\min}^{+}} \cdot \log\left(\frac{Cd^3H^4\log(4dKH/p)}{(\text{gap}_{\min}^{+})^2}\right).
\end{aligned}$$

Let $p = \frac{1}{4H^2K(K+1)}$, we can rewrite the inequality above as:

$$\mathbb{E}[\text{Regret}(K)] \leqslant 1 + \frac{Cd^3H^5\log(16dK^2(K+1)H^3)}{\text{gap}_{\min}^{+}}\iota,$$

where $\iota = \log\left(\frac{Cd^3H^4\log(4dKH/p)}{(\text{gap}_{\min}^{+})^2}\right)$ is a logarithmic term. Till now, our main theorem for centralized setting is proved. The independent version is very similar. After combining Equation (21) with the counting number upper bound (Equation (34)), we can get exactly the same result: for any sequence of policies $\mu := \{\mu^k\}_{k\in[K]}$, we have:

$$\mathbb{E}[\text{Regret}_{\mu}(K)] \leqslant 1 + \frac{Cd^3H^5\log(16dK^2(K+1)H^3)}{\text{gap}_{\min}^{+}}\iota,$$

where $\iota = \log\left(\frac{Cd^3H^4\log(4dKH/p)}{(\text{gap}_{\min}^{+})^2}\right)$ is a logarithmic term.

E The Pseudo-code of LSVI-2TBSG

In this section, we introduce the formal pseudo-code of LSVI-2TBSG algorithm in both centralized setting and independent setting.

Algorithm 3 Optimistic Nash Q-learning on two-player Turn-based Stochastic Games with Linear Function Expression (Centralized)

Initialize: Let $\overline{Q}_h^1(s,a) \leftarrow 2H, \underline{Q}_h^1(s,a) \leftarrow 0$, and $\overline{V}_h^1(s,a) \leftarrow 2H, \underline{V}_h^1(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [2H]$. $\lambda \leftarrow 1, \beta \leftarrow c_\beta \cdot dH\sqrt{\iota}$ where $c_\beta = 160$ is an absolute constant and $\iota = \log(2dT/p) = \log(4dKH/p)$.

for episode $k \in [K]$ **do**
  Observe the initial state s_1^k.
  for step $h = 2H, 2H-1, \ldots, 1$ **do**
    $\Lambda_h^k = \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$
    $\overline{w}_h^k = (\Lambda_h^k)^{-1} \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \left[r_h(s_h^i, a_h^i) + \overline{V}_{h+1}^k(s_{h+1}^i) \right]$
    $\underline{w}_h^k = (\Lambda_h^k)^{-1} \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \left[r_h(s_h^i, a_h^i) + \underline{V}_{h+1}^k(s_{h+1}^i) \right]$
    $\overline{Q}_h^k(s,a) \leftarrow \min\left(2H, \phi(s,a)^\top \overline{w}_h^k + \beta\sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)}\right)$
    $\underline{Q}_h^k(s,a) \leftarrow \max\left(0, \phi(s,a)^\top \underline{w}_h^k - \beta\sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)}\right)$
  end for
  for step $h = 1, 2, \ldots, 2H$ **do**
    If h is an odd number, take action $a_h^k \leftarrow \arg\max_a \overline{Q}_h^k(s_h^k, a)$, otherwise, take action $a_h^k \leftarrow \arg\min_a \underline{Q}_h^k(s_h^k, a)$. Then, we receive the next state s_{h+1}^k.
    $\overline{V}_h^k(s_h^k) \leftarrow \overline{Q}_h^k(s_h^k, a_h^k)$, $\underline{V}_h^k(s_h^k) \leftarrow \underline{Q}_h^k(s_h^k, a_h^k)$.
  end for
end for

Algorithm 4 Optimistic Nash Q-learning on two-player Turn-based Stochastic Games with Linear Function Expression (Independent)

Initialize: Let $\overline{Q}_h^1(s,a) \leftarrow 2H$, and $\overline{V}_h^1(s,a) \leftarrow 2H$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [2H]$. $\lambda \leftarrow 1, \beta \leftarrow c_\beta \cdot dH\sqrt{\iota}$ where $c_\beta = 160$ is an absolute constant and $\iota = \log(2dT/p) = \log(4dKH/p)$.

for episode $k \in [K]$ **do**
  Observe the initial state s_1^k.
  for step $h = 2H, 2H-1, \ldots, 1$ **do**
    $\Lambda_h^k = \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$
    $w_h^k = (\Lambda_h^k)^{-1} \sum_{i=1}^{k-1} \phi(s_h^i, a_h^i) \left[r_h(s_h^i, a_h^i) + \overline{V}_{h+1}^k(s_{h+1}^i) \right]$
    $\overline{Q}_h^k(s,a) \leftarrow \min\left(2H, \phi(s,a)^\top w_h^k + \beta\sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)}\right)$
  end for
  for step $h = 1, 2, \ldots, 2H$ **do**
    If h is an odd number, take action $a_h^k \leftarrow \arg\max_a \overline{Q}_h^k(s_h^k, a)$, otherwise, let the min-player choose his action a_h^k. Then, we receive the next state s_{h+1}^k.
    $\overline{V}_h^k(s_h^k) \leftarrow \overline{Q}_h^k(s_h^k, a_h^k)$.
  end for
end for