A Bayesian Learning Algorithm for Unknown Zero-sum Stochastic Games with an Arbitrary Opponent

Mehdi Jafarnia Jahromi
Google DeepMind

Rahul Jain
USC and Google Research

Ashutosh Nayyar
University of Southern California

Abstract

In this paper, we propose Posterior Sampling Reinforcement Learning for Zero-sum Stochastic Games (PSRL−ZSG), the first online learning algorithm that achieves Bayesian regret bound of $\widetilde{\mathcal{O}}(HS\sqrt{AT})$ in the infinite-horizon zero-sum stochastic games with average-reward criterion. Here H is an upper bound on the span of the bias function, S is the number of states, A is the number of joint actions and T is the horizon. We consider the online setting where the opponent can not be controlled and can take any arbitrary time-adaptive history-dependent strategy. Our regret bound improves on the best existing regret bound of $\widetilde{\mathcal{O}}(\sqrt[3]{DS^2AT^2})$ by Wei et al. (2017) under the same assumption and matches the theoretical lower bound in T.

1 INTRODUCTION

Recent advances in playing the game of Go (Silver et al., 2017) and Starcraft (Vinyals et al., 2019) have proved the capability of *self-play* in achieving super-human performance in competitive reinforcement learning (competitive RL) (Crandall and Goodrich, 2005), a special case of multi-agent RL where each player tries to maximize its own reward. These self-play algorithms are able to learn through repeatedly playing against themselves and update their policy based on the observed trajectory in the absence of human supervision. Despite the empirical success, the theoretical understanding of these algorithms is limited and is significantly more challenging than the single-agent RL due to its multi-agent nature.

Self-play can be considered as a special case of offline competitive RL where the learning algorithm controls both the agent and the opponent during the learning process (Bai and Jin, 2020; Bai et al., 2020). In the more general and sophisticated online learning case, the opponent can take arbitrary history-dependent strategies and the agent has no control on the opponent during the learning process (Wei et al., 2017; Xie et al., 2020; Tian et al., 2021).

In this paper, we consider the online learning setting where the agent learns against an arbitrary opponent who can follow a *time-variant history-dependent policy and can switch its policy at any time*. We consider infinite-horizon two-player zero-sum stochastic games (SGs) with the average-reward criterion. At each time, both players determine their actions simultaneously upon observing the state. The reward and the probability distribution of the next state is then determined by the chosen actions and the current state. The players' payoffs sum to zero, i.e., the reward of one player (agent) is exactly the loss of the other player (opponent). The agent's goal is to maximize its cumulative reward while the opponent tries to minimize the total loss. The problem of designing learning algorithms that can learn against arbitrary opponents is a significant open issue. There is extensive literature on designing and analyzing algorithms that learn against opponents in such a manner that they together converge to an equilibrium of the underlying game. In such cases however, the opponent is not free to choose any learning or non-learning strategy that they want, a significant limitation in their practical use.

We propose *Posterior Sampling Reinforcement Learning algorithm for Zero-sum Stochastic Games* (PSRL−ZSG), a learning algorithm that achieves $\widetilde{\mathcal{O}}(HS\sqrt{AT})$ Bayesian regret bound. Here H is an upper bound on the bias-span, S is the number of states, A is the size of all possible action pairs for both players, T is the horizon, and $\widetilde{\mathcal{O}}$ hides logarithmic factors. The best existing result in this setting is achieved by UCSG algorithm (Wei et al., 2017) which obtains a regret bound of $\widetilde{\mathcal{O}}(\sqrt[3]{DS^2AT^2})$ where $D \geq H$ is the diameter of the SG. As stochastic games generalize Markov Decision Processes (MDPs), our regret bound is optimal (except for logarithmic factors) in T due to the lower bound provided by Jaksch et al. (2010).

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Related Literature

SG was first formulated by Shapley (1953). A large body of work focuses on finding the Nash equilibria in SGs with known transition kernel (Littman, 2001; Hu and Wellman, 2003; Hansen et al., 2013), or learning with a generative model (Jia et al., 2019; Sidford et al., 2020; Zhang et al., 2020) to simulate the transition for an arbitrary state-action pair. In these cases no exploration is needed.

There is a long line of research on exploration and regret analysis in single-agent RL (see e.g. Jaksch et al. (2010); Osband et al. (2013), Gopalan and Mannor (2015); Azar et al. (2017); Ouyang et al. (2017); Jin et al. (2018); Zhang and Ji (2019); Zanette and Brunskill (2019); Wei et al. (2020, 2021); Chen et al. (2021a); Jafarnia-Jahromi et al. (2021b,a) and references therein). Extending these results to the SGs is non-trivial since the actions of the opponent also affect the state transition and can not be controlled by the agent. We review the literature on exploration in SGs and refer the interested reader to Zhang et al. (2021); Yang and Wang (2020) for an extensive literature review on multi-agent RL in various settings.

Stochastic Games. A few recent works use self-play as a method to learn stochastic games (Bai and Jin, 2020; Bai et al., 2020; Liu et al., 2021; Chen et al., 2021b). However, self-play requires controlling both the agent and the opponent and cannot be applied in the online setting where the agent plays against an arbitrary opponent. All of these works consider the setting of finite-horizon SG where the interaction of the players and the environment terminates after a fixed number of steps.

In the online setting where the opponent is *arbitrary*, Xie et al. (2020); Jin et al. (2021) achieve a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ in the finite-horizon SGs with linear and general function approximation, respectively. However, in the applications where the interaction between the players and the environment is non-stopping (e.g., stock trading), the infinite-horizon SG is more suitable. Lack of a fixed horizon in this setting makes the problem more challenging. This is since the backward induction, a technique that is widely used in the finite-horizon, is not applicable in the infinite-horizon setting. A recent paper on posterior sampling-based approaches to finite-horizon stochastic games is Zhou et al. (2020).

In the infinite-horizon setting, the primary work of Brafman and Tennenholtz (2002) who proposed R-max algorithm does not consider regret. A special case of online learning in general-sum games is studied by DiGiovanni and Tewari (2021) where the opponent is allowed to switch its stationary policy a limited number of times. They achieve a regret bound of $\widetilde{\mathcal{O}}(\ell + \sqrt{\ell T})$ via posterior sampling, where ℓ is the number of switches. Their result is not directly comparable to ours because their definition of regret is different. Moreover, they assume the transition kernel is known and the opponent adopts stationary policies. To the best of our knowledge, the only existing algorithm that considers online learning against an arbitrary opponent in the infinite-horizon average-reward SG is `UCSG` (Wei et al., 2017).

Comparison with `UCSG` (Wei et al., 2017). Our work is closely related to `UCSG`, however clear distinctions exist in the result, the algorithm, and the technical contribution:

- `UCSG` achieves a regret bound of $\widetilde{\mathcal{O}}(\sqrt[3]{DS^2AT^2})$ under the finite-diameter assumption (i.e., for any two states and every stationary randomized policy of the opponent, there *exists* a stationary randomized policy for the agent to move from one state to the other in finite expected time). Under the much stronger ergodicity assumption (i.e., for any two states and *every* stationary randomized policy of the agent and the opponent, it is possible to move from one state to the other in finite expected time), `UCSG` obtains a regret bound of $\widetilde{\mathcal{O}}(DS\sqrt{AT})$. Note that the ergodicity assumption greatly alleviates the challenge in exploration. Our algorithm significantly improves this result and achieves a regret bound of $\widetilde{\mathcal{O}}(HS\sqrt{AT})$ under the finite-diameter assumption.

- `UCSG` is an optimism-based algorithm inspired by Jaksch et al. (2010) and requires the complicated maximin extended value iteration. Our algorithm, however, is the first posterior sampling-based algorithm in SGs, leveraging the ideas of Ouyang et al. (2017) in MDPs, and is much simpler both in the algorithm and the analysis. Note that considering randomized policies in SGs (compared to MDPs) brings some challenges in applying the concentration bounds because of the continuous space of randomized policies. However, we handle this by simply using the tower property of conditional expectation which allows us to replace the continuous space of randomized policies with the finite space of actions.

- From the analysis perspective, under the finite-diameter assumption, `UCSG` uses a sequence of finite-horizon SGs to approximate the average-reward SG and that leads to the sub-optimal regret bound of $\mathcal{O}(T^{2/3})$. Our analysis avoids the finite-horizon approximation by directly using the Bellman equation in the infinite-horizon SG and achieves near-optimal regret bound.

We note that the main challenge in online learning in a Stochastic Game (SG) is the opponent's non-stationarity and uncontrollability. Wei et al. (2017) developed a technique to replace the opponent's non-stationary policy with

a stationary one in their analysis leading to a very complicated analysis and sub-optimal regret bound. We significantly simplify the analysis and improve the final regret bound with a novel technique in which we replace the opponent's policy with arbitrary distribution over actions.

2 PRELIMINARIES

Let $M = (\mathcal{S}, \mathcal{A}, r, \theta)$ be a stochastic zero-sum game where $\mathcal{S}$ is the state space, $\mathcal{A} = \mathcal{A}^1 \times \mathcal{A}^2$ is the joint action space, $r : \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \to [-1, 0]$ is the reward function and $\theta : \mathcal{S} \times \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2$ represents the transition kernel such that $\theta(s'|s, a^1, a^2) = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t^1 = a^1, a_t^2 = a^2)$ where $s_t \in \mathcal{S}, a_t^1 \in \mathcal{A}^1, a_t^2 \in \mathcal{A}^2$ are the state, the agent and the opponent's actions at time $t = 1, 2, 3, \cdots$, respectively. We assume that $\mathcal{S}, \mathcal{A}$ are finite sets with size $S = |\mathcal{S}|, A = |\mathcal{A}|$.

The game starts at some initial state s_1. At time $t = 1, 2, 3, \cdots$, the players observe state s_t and take actions a_t^1, a_t^2. The agent (maximizer) receives reward $r(s_t, a_t^1, a_t^2)$ from the opponent (minimizer). Then, the state evolves to s_{t+1} according to the probability distribution $\theta(\cdot|s_t, a_t^1, a_t^2)$. The goal of the agent is to maximize its cumulative reward while the opponent tries to minimize it. For the ease of notation, we denote $a := (a^1, a^2)$ and $a_t := (a_t^1, a_t^2)$ and accordingly $r(s_t, a_t^1, a_t^2), \theta(\cdot|s_t, a_t^1, a_t^2)$ will be denoted by $r(s_t, a_t)$ and $\theta(\cdot|s_t, a_t)$, respectively.

The players' actions are assumed to depend on the history. Namely, denote by π_t^1 (resp. π_t^2) the mappings from the history $h_t = (s_1, a_1, \cdots, s_{t-1}, a_{t-1}, s_t)$ to the probability distributions over $\mathcal{A}_1$ (resp. $\mathcal{A}_2$). Let $\pi^1 := (\pi_1^1, \pi_2^1, \cdots)$ (resp. $\pi^2 := (\pi_1^2, \pi_2^2, \cdots)$) be the sequence of history-dependent randomized policies whose class is denoted by Π^{HR}. In the case that π_t^1 (resp. π_t^2) is independent of time (stationary randomized policies), we remove the subscript t and with abuse of notation denote $\pi^1 := (\pi^1, \pi^1, \cdots)$ (resp. $\pi^2 := (\pi^2, \pi^2, \cdots)$). The class of stationary randomized policies is denoted by Π^{SR}.

For the ease of presentation, we introduce a few notations. Let $A^1 = |\mathcal{A}^1|$, $A^2 = |\mathcal{A}^2|$ denote the size of the action spaces. For an integer $k \geq 1$, denote by Δ_k the probability simplex of dimension k. Let $q^1 \in \Delta_{A^1}$ and $q^2 \in \Delta_{A^2}$. With abuse of notation, let $r(s, q^1, q^2) := \mathbb{E}_{a^1 \sim q^1, a^2 \sim q^2}[r(s, a^1, a^2)]$ and $\theta(s'|s, q^1, q^2) := \mathbb{E}_{a^1 \sim q^1, a^2 \sim q^2}[\theta(s'|s, a^1, a^2)]$.

To achieve a low regret algorithm, it is necessary to assume that all the states are accessible by the agent under some policy. In the special case of MDPs, this is stated by the notion of "weakly communication" (or "finite diameter" (Jaksch et al., 2010)) and is known to be the minimal assumption to achieve sub-linear regret (Bartlett and Tewari, 2009). The following assumption generalizes this notion to the stochastic games.

Assumption 2.1. (Finite Diameter) There exists $D \geq 0$ such that for any stationary randomized policy $\pi^2 \in \Pi^{\text{SR}}$ of the opponent and any $s, s' \in \mathcal{S} \times \mathcal{S}$, there exists a stationary randomized policy $\pi^1 \in \Pi^{\text{SR}}$ of the agent, such that the expected time of reaching s' starting from s under policy $\pi = (\pi^1, \pi^2)$ does not exceed D, i.e.,

$$\max_{s,s'} \max_{\pi^2 \in \Pi^{\text{SR}}} \min_{\pi^1 \in \Pi^{\text{SR}}} T_{s \to s'}^{\pi} \leq D,$$

where $T_{s \to s'}^{\pi}$ is the expected time of reaching s' starting from s under policy $\pi = (\pi^1, \pi^2)$.

This assumption was first introduced by Federgruen (1978) and is essential to achieve low regret algorithms in the adversarial setting (Wei et al., 2017). To see this, suppose that the opponent has a way to lock the agent in a "bad" state. In the initial stages of the game when the agent has limited environment knowledge, it may not be possible to avoid such a state and linear regret is unavoidable. This assumption states that regardless of the strategy used by the opponent, the agent has a way to recover from such bad states.

For a zero-sum matrix game with matrix G of size $m \times n$, the game value is denoted by $\text{val}(G) = \max_{p \in \Delta_m} \min_{q \in \Delta_n} p^T G q = \min_{q \in \Delta_n} \max_{p \in \Delta_m} p^T G q$. Moreover, the Nash equilibrium $p^* \in \Delta_m, q^* \in \Delta_n$ always exists (Nash et al., 1950). For SGs, under Assumption 2.1, Federgruen (1978); Wei et al. (2017) prove that there exist unique $J(\theta) \in \mathbb{R}$ and unique (upto an additive constant) function $v(\cdot, \theta) : \mathcal{S} \to \mathbb{R}$ that satisfy the Bellman equation, i.e., for all $s \in \mathcal{S}$,

$$J(\theta) + v(s, \theta) = \text{val}\left\{ r(s, \cdot, \cdot) + \sum_{s'} \theta(s'|s, \cdot, \cdot) v(s', \theta) \right\}. \tag{1}$$

In particular, the Nash equilibrium of the right hand side for each $s \in \mathcal{S}$ yields maximin stationary policies $\pi^* = (\pi^{1*}, \pi^{2*})$ such that

$$J(\theta) + v(s, \theta) = \max_{q^1 \in \Delta_{A^1}} \Big\{ r(s, q^1, \pi^{2*}(\cdot|s)) + \sum_{s'} \theta(s'|s, q^1, \pi^{2*}(\cdot|s)) v(s', \theta) \Big\}, \tag{2}$$

$$J(\theta) + v(s, \theta) = \min_{q^2 \in \Delta_{A^2}} \Big\{ r(s, \pi^{1*}(\cdot|s), q^2) + \sum_{s'} \theta(s'|s, \pi^{1*}(\cdot|s), q^2) v(s', \theta) \Big\}. \tag{3}$$

Moreover, $J(\theta)$ is the maximin average reward obtained by the agent and is independent of the initial state s_1, i.e.,

$$J(\theta) = \sup_{\pi^1 \in \Pi^{\text{HR}}} \inf_{\pi^2 \in \Pi^{\text{HR}}} \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t) | s_1 = s \right],$$

where $a_t = (a_t^1, a_t^2)$ and $a_t^1 \sim \pi_t^1(\cdot|h_t)$ and $a_t^2 \sim \pi_t^2(\cdot|h_t)$. Note that $J(\theta) \in [-1, 0]$ because the range of the reward function is $[-1, 0]$. Define the *span* of the stochastic game with transition kernel θ as the span of the corresponding value function v, i.e., $\mathrm{sp}(\theta) := \max_s v(s, \theta) - \min_s v(s, \theta)$. We restrict our attention to stochastic games whose transition kernel θ satisfies Assumption 2.1 and $\mathrm{sp}(\theta) \leq H$ where H is a known scalar. This constant is not used explicitly in the algorithm we propose but is implicit since all transition kernels we allow have bias-span bounded by H. Let Ω_* denote the set of all such θ. Moreover, observe that if v satisfies the Bellman equation, $v + c$ also satisfies the Bellman equation for any scalar c. Thus, without loss of generality, we can assume that $0 \leq v(s, \theta) \leq H$ for all $s \in \mathcal{S}$ and $\theta \in \Omega_*$.

Stationary Randomized Opponent. We first consider the special case where the opponent follows a fixed unknown stationary randomized policy π^2. In that case, the agent can consider the opponent as part of the environment and define a new environment with reward and transition kernel

$$r^{\pi^2}(s, a^1) := r(s, a^1, \pi^2(s)),$$
$$\theta^{\pi^2}(s'|s, a^1) := \theta(s'|s, a^1, \pi^2(s)).$$

Since the new environment is stationary, the agent can use any standard single-agent RL algorithm. For example, applying TSDE algorithm (Ouyang et al., 2017) yields a regret bound of $\widetilde{\mathcal{O}}(DS\sqrt{A^1T})$.[1] The rest of the paper considers the more general case where the opponent can take any time-adaptive randomized policy.

Time-adaptive Randomized Opponent. The focus of this paper is on the case where the agent plays a stochastic game $(\mathcal{S}, \mathcal{A}, r, \theta_*)$ against an opponent who can take time-adaptive policies. We assume that the opponent knows the history of states and actions and can play time-adaptive history-dependent policies. Recall that the state of such policies is denoted by Π^{HR}. Considering the opponent as part of the environment in this case results in a time-varying environment and, therefore, standard single-agent no-regret algorithms are not applicable. $\mathcal{S}, \mathcal{A}$ and r are completely known to the agent. However, the transition kernel θ_* is unknown. In the beginning of the game, θ_* is drawn from an initial distribution μ_1 and is then fixed. We assume that the support of μ_1 is a subset of Ω_*. The performance of the agent is then measured with the notion of regret defined as

$$R_T := \sup_{\pi^2 \in \Pi^{\mathrm{HR}}} \mathbb{E}\left[\sum_{t=1}^{T} (J(\theta_*) - r(s_t, a_t))\right],$$

where $a_t^2 \sim \pi_t^2(\cdot|h_t)$. Here the expectation is with respect to the prior distribution μ_1, randomized algorithm and the randomness in the state transition. Note that the regret guarantee is against an arbitrary opponent who can change its policy at each time step and has the perfect knowledge of the history of the states and actions. The only hidden information from the opponent is the realization of the agent's current action (which will be revealed after both players have chosen their actions). We note that self-play and the case when the agent and the opponent use the same learning algorithm are two special cases of the scenario considered here.

3 POSTERIOR SAMPLING FOR STOCHASTIC GAMES

In this section, we propose Posterior Sampling algorithm for Zero-sum SGs (PSRL-ZSG). The agent maintains the posterior distribution μ_t on parameter θ_*. More precisely, the learning algorithm receives an initial distribution μ_1 as the input and updates the posterior distribution upon observing the new state according to

$$\mu_{t+1}(d\theta) \propto \theta(s_{t+1}|s_t, a_t)\mu_t(d\theta). \tag{4}$$

PSRL-ZSG proceeds in episodes. Let t_k, T_k denote the start time and the length of episode k, respectively. In the beginning of each episode, the agent draws a sample of the transition kernel from the posterior distribution μ_{t_k}. The maximin strategy is then derived for the sampled transition kernel according to (1) and used by the agent during the episode. Let $N_t(s, a)$ be the number of visits to state-action pair $(s, a) = (s, a^1, a^2)$ before time t, i.e.,

$$N_t(s, a) = \sum_{\tau=1}^{t-1} \mathbb{1}(s_\tau = s, a_\tau = a).$$

As described in Algorithm 1, a new episode starts if $t > t_k + T_{k-1}$ or $N_t(s, a) > 2N_{t_k}(s, a)$ for some (s, a). The first criterion, $t > t_k + T_{k-1}$, states that the length of the episode grows at most by 1 if the other criterion is not triggered. This ensures that $T_k \leq T_{k-1} + 1$ for all k. The second criterion is triggered if the number of visits to a state-action pair is doubled. These stopping criteria balance the trade-off between exploration and exploitation. In the beginning of the game, the episodes are short to motivate exploration since the agent is uncertain about the underlying environment. As the game proceeds, the episodes grow to exploit the information gathered about the environment. These stopping criteria are the same as those used in MDPs (Ouyang et al., 2017).

Algorithm 1 can achieve regret bound of $\widetilde{\mathcal{O}}(HS\sqrt{AT})$. This result improves upon the previous best known result of UCSG algorithm which achieves $\widetilde{\mathcal{O}}(\sqrt[3]{DS^2AT^2})$ under the same assumption (Wei et al., 2017).

[1] The original bound in Ouyang et al. (2017) is $\widetilde{\mathcal{O}}(HS\sqrt{A^1T})$ where H is an upper bound on the span of the relative value function. Assumption 2.1 implies that the diameter and thus the span of the relative value function of the induced MDP is upper bounded by D.

Algorithm 1 PSRL-ZSG

Input: μ_1
Initialization: $t \leftarrow 1, t_1 \leftarrow 0$
for *episodes* $k = 1, 2, \cdots$ **do**
  $T_{k-1} \leftarrow t - t_k$
  $t_k \leftarrow t$
  Generate $\theta_k \sim \mu_{t_k}$ and compute $\pi_k^1(\cdot)$ using (1)
  while $t \leq t_k + T_{k-1}$ *and* $N_t(s,a) \leq 2N_{t_k}(s,a)$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
    Choose action $a_t^1 \sim \pi_k^1(\cdot|s_t)$ and observe a_t^2, s_{t+1}
    Update μ_{t+1} according to (4)
    $t \leftarrow t + 1$

Theorem 3.1. *Under Assumption 2.1, Algorithm 1 can achieve regret bound of*

$$R_T \leq (H+1)\sqrt{2SAT\log T} + H + H\left(SA + 2\sqrt{SAT}\right)\sqrt{224S\log(2AT)}. \quad (5)$$

4 ANALYSIS

In this section, we provide the proof of Theorem 3.1. A central observation in our analysis is that in the beginning of each episode, θ_* and θ_k are identically distributed conditioned on the history. This key property of posterior sampling relates quantities that depend on the unknown θ_* to those of the sampled θ_k which is fully observed by the agent. Posterior sampling ensures that if t_k is a stopping time, for any measurable function f and any h_{t_k}-measurable random variable X, $\mathbb{E}[f(\theta_*, X)|h_{t_k}] = \mathbb{E}[f(\theta_k, X)|h_{t_k}]$ (Ouyang et al., 2017; Osband et al., 2013).

The key challenge in the analysis of stochastic games is that the opponent is also making decisions. If the opponent follows a fixed stationary policy, it can be considered as part of the environment and thus the SG reduces to an MDP. However, in the case that the opponent uses a dynamic history-dependent policy during the learning phase of the agent, this reduction is not possible. The key lemma in our analysis is Lemma 4.2 which overcomes this difficulty through the Bellman equation for the SG.

4.1 Proof of Theorem 3.1

Let $K_T := \max\{k : t_k \leq T\}$ be the number of episodes until time T and define $t_{K_T+1} = T+1$. Recall that $R_T = \sup_{\pi^2 \in \Pi^{\text{HR}}} R_T(\pi^2)$ where

$$R_T(\pi^2) = \mathbb{E}\left[TJ(\theta_*) - \sum_{t=1}^{T} r(s_t, a_t)\right]. \quad (6)$$

Let $\pi^2 \in \Pi^{\text{HR}}$ be an arbitrary history-dependent randomized strategy followed by the opponent. We start by decomposing the regret into two terms

$$\begin{aligned} R_T(\pi^2) &= \mathbb{E}\left[TJ(\theta_*) - \sum_{t=1}^{T} r(s_t, a_t)\right] \\ &= \mathbb{E}\left[TJ(\theta_*) - \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} J(\theta_k)\right] \\ &\quad + \mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} (J(\theta_k) - r(s_t, a_t))\right]. \end{aligned} \quad (7)$$

Lemma 4.1 uses the property of posterior sampling to bound the first term. The second term is handled by combining the Bellman equation, concentration inequalities and the property of posterior sampling as detailed in Lemma 4.2. Finally, Lemma 4.3 bounds the number of episodes and completes the proof.

Lemma 4.1. *The first term of* (7) *can be bounded by*

$$\mathbb{E}\left[TJ(\theta_*) - \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} J(\theta_k)\right] \leq \mathbb{E}[K_T]$$

Proof.

$$\begin{aligned} \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} J(\theta_k) &= \sum_{k=1}^{K_T} T_k J(\theta_k) = \sum_{k=1}^{\infty} \mathbb{1}(t_k \leq T) T_k J(\theta_k) \\ &\geq \sum_{k=1}^{\infty} \mathbb{1}(t_k \leq T)(T_{k-1}+1) J(\theta_k) \end{aligned} \quad (8)$$

where the last inequality is by the fact that $J(\theta_k) \leq 0$ and $T_k \leq T_{k-1} + 1$ due to the first stopping criterion. Now, note that t_k is a stopping time and $\mathbb{1}(t_k \leq T)$ and T_{k-1} are h_{t_k}-measurable random variables. Thus, by the property of posterior sampling and monotone convergence theorem,

$$\begin{aligned} &\mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}(t_k \leq T)(T_{k-1}+1) J(\theta_k) \,\middle|\, h_{t_k}\right] \\ &= \sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{1}(t_k \leq T)(T_{k-1}+1) J(\theta_k) \mid h_{t_k}\right] \\ &= \sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{1}(t_k \leq T)(T_{k-1}+1) J(\theta_*) \mid h_{t_k}\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}(t_k \leq T)(T_{k-1}+1) J(\theta_*) \,\middle|\, h_{t_k}\right] \\ &\geq \mathbb{E}\left[\sum_{k=1}^{K_T} (T_{k-1}+1) J(\theta_*) \,\middle|\, h_{t_k}\right]. \end{aligned}$$

Taking another expectation from both sides and using the tower property, we have

$$\mathbb{E}\left[\sum_{k=1}^{\infty}\mathbb{1}(t_k \le T)(T_{k-1}+1)J(\theta_k)\right] \ge \mathbb{E}\left[\sum_{k=1}^{K_T}(T_{k-1}+1)J(\theta_*)\right].$$

Replacing this in (8) implies that

$$\mathbb{E}\left[TJ(\theta_*) - \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} J(\theta_k)\right] \le \mathbb{E}\left[(T - \sum_{k=1}^{K_T} T_{k-1})J(\theta_*)\right] - \mathbb{E}[K_T J(\theta_*)] \le \mathbb{E}[K_T].$$

The last inequality is by the fact that $T - \sum_{k=1}^{K_T} T_{k-1} \le 0$ and $J(\theta_*) \in [-1, 0]$. □

Lemma 4.2. *The second term of* (7) *can be bounded by*

$$\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}(J(\theta_k) - r(s_t, a_t))\right] \le H\mathbb{E}[K_T] + H + \sqrt{224S\log(2AT)}(HSA + 2H\sqrt{SAT}).$$

Proof. The policy π_k^1 used by the agent at episode k is the solution of the Nash equilibrium in (1). Thus, for $t_k \le t \le t_{k+1} - 1$ and any $s \in \mathcal{S}$, (3) implies that

$$J(\theta_k) + v(s, \theta_k) \le r(s, \pi_k^1(\cdot|s), q^2) + \sum_{s'}\theta_k(s'|s, \pi_k^1(\cdot|s), q^2)v(s', \theta_k),$$

for any distribution $q^2 \in \Delta_{A^2}$. Let $\pi^2 = (\pi_1^2, \pi_2^2, \cdots) \in \Pi^{\text{HR}}$ be an arbitrary history-dependent randomized strategy for the opponent. Note that for any $t \ge 1$, π_t^2 is h_t-measurable. Replacing s by s_t and q^2 by $\pi_t^2(\cdot|h_t)$ implies that

$$J(\theta_k) - r(s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t)) \le \sum_{s'}\theta_k(s'|s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t))v(s', \theta_k) - v(s_t, \theta_k).$$

Adding and subtracting $v(s_{t+1}, \theta_k)$ to the right hand side and summing over time steps within episode k implies that

$$\begin{aligned}\sum_{t=t_k}^{t_{k+1}-1}\big(J(\theta_k) - r(s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t))\big) &\le \sum_{t=t_k}^{t_{k+1}-1}\left(\sum_{s'}\theta_k(s'|s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t))v(s', \theta_k) - v(s_{t+1}, \theta_k)\right) \\ &\quad + \sum_{t=t_k}^{t_{k+1}-1}(v(s_{t+1}, \theta_k) - v(s_t, \theta_k)). \end{aligned} \tag{9}$$

The second term on the right hand side of (9) telescopes and can be bounded as

$$\sum_{t=t_k}^{t_{k+1}-1}(v(s_{t+1}, \theta_k) - v(s_t, \theta_k)) = v(s_{t_{k+1}}, \theta_k) - v(s_{t_k}, \theta_k) \le H, \tag{10}$$

where the last inequality is by the fact that θ_k is chosen from the posterior distribution whose support is a subset of Ω_*. Substituting (10) in (9), summing over episodes, and taking expectation implies that

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}(J(\theta_k) - r(s_t, a_t))\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\big(J(\theta_k) - r(s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t))\big)\right] \\ &\le H\mathbb{E}[K_T] + \mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \\ &\sum_{s'}\theta_k(s'|s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t))v(s', \theta_k) - v(s_{t+1}, \theta_k)\Bigg].\end{aligned}$$

We proceed to bound the last term on the right hand side of the above inequality. Before proceeding note that if $k(t)$ denotes the episode at time t, a random variable. Then, for any $t \ge 1$ and $s' \in \mathcal{S}$,

$$\mathbb{E}\Big[\theta_{k(t)}(s'|s_t, a_t^1, a_t^2)\big|h_t, \theta_{k(t)}\Big] = \theta_{k(t)}\Big(s'\big|s_t, \pi_{k(t)}^1(\cdot|s_t), \pi_t^2(\cdot|h_t)\Big), \qquad (*)$$

because $a_t^1 \sim \pi_{k(t)}^1(\cdot|s_t)$ and $a_t^2 \sim \pi_t^2(\cdot|h_t)$. Now,

$$\begin{aligned}&\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \\ &\sum_{s'}\theta_k(s'|s_t, \pi_k^1(\cdot|s_t), \pi_t^2(\cdot|h_t))v(s', \theta_k) - v(s_{t+1}, \theta_k)\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \\ &\sum_{s'}\theta_k(s'|s_t, a_t^1, a_t^2)v(s', \theta_k) - v(s_{t+1}, \theta_k)\Bigg] = \\ &\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sum_{s'}[\theta_k(s'|s_t, a_t) - \theta_*(s'|s_t, a_t)]\,v(s', \theta_k)\right] \\ &\le H\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sum_{s'}\Big|\,\theta_k(s'|s_t, a_t) - \theta_*(s'|s_t, a_t)\,\Big|\right]\end{aligned} \tag{11}$$

To bound the inner summation, similar to Ouyang et al. (2017); Jaksch et al. (2010), we define a confidence set

$\mathcal{C}_k$ around the empirical transition kernel $\hat{\theta}_k(s'|s,a) := \frac{N_{t_k}(s',s,a)}{N_{t_k}(s,a)}$. Here $N_{t_k}(s',s,a) := \sum_{t=1}^{t_k-1} \mathbb{1}(s_t = s, a_t = a, s_{t+1} = s')$ is the number of visits to state-action pair (s,a) whose next state is s'. The confidence set $\mathcal{C}_k$ is defined as $\mathcal{C}_k :=$

$$\{\theta : \sum_{s'} |\theta(s'|s,a) - \hat{\theta}_k(s'|s,a)| \le b_k(s,a) \quad \forall s,a,s'\},$$

where $b_k(s,a) := \sqrt{\frac{14S\log(2At_kT)}{\max\{1,N_{t_k}(s,a)\}}}$. Weissman et al. (2003) shows that the true transition kernel θ_* belongs to $\mathcal{C}_k$ with high probability. We use this fact to show concentration of $\hat{\theta}_k$ around θ_*. Concentration of $\hat{\theta}_k$ around θ_k is then followed by the property of posterior sampling. More precisely, we can write

$$\begin{aligned}
&\sum_{s'} |\theta_k(s'|s_t,a_t) - \theta_*(s'|s_t,a_t)| \\
&\le \sum_{s'} |\theta_k(s'|s_t,a_t) - \hat{\theta}_k(s'|s_t,a_t)| \\
&\quad + \sum_{s'} |\theta_*(s'|s_t,a_t) - \hat{\theta}_k(s'|s_t,a_t)| \\
&\le 2b_k(s_t,a_t) + 2\left(\mathbb{1}(\theta_k \notin \mathcal{C}_k) + \mathbb{1}(\theta_* \notin \mathcal{C}_k)\right).
\end{aligned}$$

Substituting the inner sum of (11) with this upper bound implies

$$\begin{aligned}
&H\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sum_{s'}\left|\theta_k(s'|s_t,a_t) - \theta_*(s'|s_t,a_t)\right|\right] \\
&\le 2H\left\{\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} b_k(s_t,a_t)\right\} \\
&\quad + 2H\mathbb{E}\left[\sum_{k=1}^{K_T} T_k\{\mathbb{1}(\theta_k \notin \mathcal{C}_k) + \mathbb{1}(\theta_* \notin \mathcal{C}_k)\}\right].
\end{aligned} \tag{12}$$

The first term on the right hand side of (12) can be bounded as

$$\begin{aligned}
\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} b_k(s_t,a_t) &= \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{14S\log(2At_kT)}{\max\{1,N_{t_k}(s_t,a_t)\}}} \\
&\le \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{28S\log(2AT^2)}{\max\{1,N_t(s_t,a_t)\}}} \\
&= \sum_{t=1}^{T}\sqrt{\frac{28S\log(2AT^2)}{\max\{1,N_t(s_t,a_t)\}}} \\
&\le \sqrt{56S\log(2AT)}(SA + 2\sqrt{SAT}),
\end{aligned} \tag{13}$$

where the first inequality is by the fact that $t_k \le T$ and $N_t(s,a) \le 2N_{t_k}(s,a)$ for all s,a and the second inequality is by the following argument:

$$\begin{aligned}
&\sum_{t=1}^{T}\sqrt{\frac{1}{\max\{1,N_t(s_t,a_t)\}}} = \sum_{t=1}^{T}\sum_{s,a}\frac{\mathbb{1}(s_t=s,a_t=a)}{\sqrt{\max\{1,N_t(s,a)\}}} \\
&= \sum_{s,a}\sum_{t=1}^{T}\frac{\mathbb{1}(s_t=s,a_t=a)}{\sqrt{\max\{1,N_t(s,a)\}}} \\
&= \sum_{s,a}\left(1 + \sum_{j=1}^{n_{T+1}(s,a)-1}\frac{1}{\sqrt{j}}\right) \\
&\le \sum_{s,a}\left(1 + 2\sqrt{N_{T+1}(s,a)}\right) = SA + 2\sum_{s,a}\sqrt{N_{T+1}(s,a)} \\
&\le SA + 2\sqrt{SA\sum_{s,a}N_{T+1}(s,a)} = SA + 2\sqrt{SAT},
\end{aligned}$$

where the last inequality is by Cauchy-Schwarz and the last equality is by the fact that $\sum_{s,a} N_{T+1}(s,a) = T$. To bound the second term on the right hand side of (12), we can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K_T} T_k\{\mathbb{1}(\theta_k \notin \mathcal{C}_k) + \mathbb{1}(\theta_* \notin \mathcal{C}_k)\}\right] \\
&\le \mathbb{E}\left[\sum_{k=1}^{\infty} T\{\mathbb{1}(\theta_k \notin \mathcal{C}_k) + \mathbb{1}(\theta_* \notin \mathcal{C}_k)\}\right] \\
&= T\sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{1}(\theta_k \notin \mathcal{C}_k) + \mathbb{1}(\theta_* \notin \mathcal{C}_k)\right] \\
&= 2T\sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{1}(\theta_* \notin \mathcal{C}_k)\right] = 2T\sum_{k=1}^{\infty}\mathbb{P}(\theta_* \notin \mathcal{C}_k),
\end{aligned}$$

where the second equality is by the property of Posterior Sampling since $\mathcal{C}_k$ is $\mathcal{F}_{t_k}$-measurable. Note that $\mathbb{P}(\theta_* \notin \mathcal{C}_k) \le \frac{1}{15Tt_k^6}$ (Lemma 17 of Jaksch et al. (2010)). Thus,

$$2T\sum_{k=1}^{\infty}\mathbb{P}(\theta_* \notin \mathcal{C}_k) = \frac{2}{15}\sum_{k=1}^{\infty}\frac{1}{t_k^6} \le \frac{2}{15}\sum_{k=1}^{\infty}\frac{1}{k^6} \le \frac{1}{2}. \tag{14}$$

Combining (13) and (14) in (12) completes the proof. □

Remark 1: We note that the bias span of v is bounded by H because we are using the bias span of v from the minimax Bellman equation (not from the Bellman equation of the player's policies.) This key observation allows us to handle an adversarial opponent and improve the bound of prior work (Wei et al., 2017) with a much simpler analysis.

It remains to bound the number of episodes. The following lemma completes the proof of Theorem 3.1.

Lemma 4.3. *The number of episodes can be bounded by*

$$K_T \le \sqrt{2SAT\log T}.$$

Proof. Define macro episodes with start times t_{m_i} as $t_{m_1} = t_1$ and for $i \ge 2$, $t_{m_i} :=$

$$\min\{t_k > t_{m_{i-1}} : n_{t_k}(s,a) > 2n_{t_{k-1}}(s,a), \text{for some } (s,a)\}.$$

Note that t_{m_i} is the start time of the ith macro episode and corresponds to the ith start time that an episode triggers with the second stopping criterion in Algorithm 1. Denote by M_T the number of macro episodes by time T and let $m_{M_T+1} = K_T + 1$.

Let $\tilde{T}_i$ be the length of the ith macro episode. We can write $\tilde{T}_i = \sum_{k=m_i}^{m_{i+1}-1} T_k$. All the episodes except the last one within a macro episode are started with the first criterion. Thus, for all $m_i \le k \le m_{i+1} - 2$, $T_k = T_{k-1} + 1$, and

$$\begin{aligned}\tilde{T}_i &= \sum_{k=m_i}^{m_{i+1}-1} T_k = T_{m_{i+1}-1} + \sum_{j=1}^{m_{i+1}-m_i-1} (T_{m_i-1} + j) \\ &\ge 1 + \sum_{j=1}^{m_{i+1}-m_i-1} (1+j) \\ &= 0.5(m_{i+1} - m_i)(m_{i+1} - m_i + 1).\end{aligned}$$

This implies that $m_{i+1} - m_i \le \sqrt{2\tilde{T}_i}$ for all $i = 1, \cdots, M_T$. Consequently,

$$\begin{aligned}K_T = m_{M_T+1} - 1 &= \sum_{i=1}^{M_T} (m_{i+1} - m_i) \le \sum_{i=1}^{M_T} \sqrt{2\tilde{T}_i} \\ &\le \sqrt{2M_T \sum_{i=1}^{M_T} \tilde{T}_i} = \sqrt{2M_T T}, \qquad (15)\end{aligned}$$

where the last inequality is by Cauchy-Schwarz and the last equality is due to $\sum_{i=1}^{M_T} \tilde{T}_i = T$. Now, it suffices to prove that $M_T \le SA \log T$. To see this, let $\mathcal{T}_{s,a}$ be the episode start times that are triggered by the second stopping criterion at state-action pair (s,a). That is,

$$\mathcal{T}_{s,a} := \{t_k \le T : n_{t_k}(s,a) > 2n_{t_{k-1}}(s,a)\}.$$

Since the number of visits to state-action pair (s,a) is doubled at each $t_k \in \mathcal{T}_{s,a}$, we claim that $|\mathcal{T}_{s,a}| \le \log n_{T+1}(s,a)$. To see this, assume by contradiction that $|\mathcal{T}_{s,a}| > \log n_{T+1}(s,a) + 1$. We can write

$$\begin{aligned}n_{t_{K_T}}(s,a) &\ge \prod_{t_k \le T, n_{t_{k-1}}(s,a) \ge 1} \frac{n_{t_k}(s,a)}{n_{t_{k-1}}(s,a)} \\ &\ge \prod_{t_k \in \mathcal{T}_{s,a}, n_{t_{k-1}}(s,a) \ge 1} \frac{n_{t_k}(s,a)}{n_{t_{k-1}}(s,a)} \\ &> \prod_{t_k \in \mathcal{T}_{s,a}, n_{t_{k-1}}(s,a) \ge 1} 2 = 2^{|\mathcal{T}_{s,a}|-1} \ge n_{T+1}(s,a),\end{aligned}$$

which is a contradiction. Here, the second inequality is by the fact that $n_t(s,a)$ is non-decreasing and the last inequality is by the definition of $\mathcal{T}_{s,a}$. Now, we can write

$$\begin{aligned}M_T = 1 + |\mathcal{T}_{s,a}| &\le 1 + \sum_{s,a} \log n_{T+1}(s,a) \\ &\le 1 + SA \log(\sum_{s,a} n_{T+1}(s,a)/SA) \\ &= 1 + SA \log(T/SA) \le SA \log T.\end{aligned}$$

Here, the second inequality is by the concavity of $\log$. Replacing this inequality in (15) completes the proof.

□

5 CONCLUSIONS

We proposed PSRL−ZSG, a posterior sampling algorithm that achieves Bayesian regret bound of $\tilde{\mathcal{O}}(HS\sqrt{AT})$ in the infinite-horizon zero-sum stochastic games with average-reward criterion. No structure is imposed on the opponent's strategy. The best existing result achieves high probability regret bound of $\tilde{\mathcal{O}}(DS\sqrt{AT})$ only under the strong ergodicity assumption. PSRL−ZSG relaxes that assumption and improves the previous best known high probability regret bound of $\tilde{\mathcal{O}}(\sqrt[3]{DS^2AT^2})$ obtained by UCSG algorithm (Wei et al., 2017) under the same finite diameter assumption. This bound is order optimal in terms of A and T. The framework and analysis developed in this paper may be useful for designing regret-optimal algorithms based on the optimism in face of uncertainty principle for zero-sum stochastic games.

Please note that in a game situation, it is very challenging to have an experimental setup from which we can draw meaningful conclusions since the opponent is free to do whatever they want. A direction for future work would be to assess the proposed algorithm in a systematic manner empirically.

References

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

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

Y. Bai, C. Jin, and T. Yu. Near-optimal reinforcement learning with self-play. In *Advances in Neural Information Processing Systems*, pages 2159–2170, 2020.

P. L. Bartlett and A. Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

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

L. Chen, M. Jafarnia-Jahromi, R. Jain, and H. Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *arXiv preprint arXiv:2106.08377*, 2021a.

Z. Chen, D. Zhou, and Q. Gu. Almost optimal algorithms for two-player markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021b.

J. W. Crandall and M. A. Goodrich. Learning to compete, compromise, and cooperate in repeated general-sum games. In *Proceedings of the 22nd international conference on machine learning*, pages 161–168, 2005.

A. DiGiovanni and A. Tewari. Thompson sampling for markov games with piecewise stationary opponent policies. In *Proceedings of the 37th Annual Conference on Uncertainty in Artificial Intelligence*, 2021.

A. Federgruen. On n-person stochastic games by denumerable state space. *Advances in Applied Probability*, 10(2): 452–471, 1978.

A. Gopalan and S. Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pages 861–898, 2015.

T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

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

M. Jafarnia-Jahromi, L. Chen, R. Jain, and H. Luo. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021a.

M. Jafarnia-Jahromi, R. Jain, and A. Nayyar. Online learning for unknown partially observable mdps. *arXiv preprint arXiv:2102.12661*, 2021b.

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

Z. Jia, L. F. Yang, and M. Wang. Feature-based q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

C. Jin, Q. Liu, and T. Yu. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021.

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

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

J. F. Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1): 48–49, 1950.

I. Osband, D. Russo, and B. Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342, 2017.

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

A. Sidford, M. Wang, L. Yang, and Y. Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR, 2020.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Y. Tian, Y. Wang, T. Yu, and S. Sra. Online learning in unknown markov games. In *International Conference on Machine Learning*, pages 10279–10288. PMLR, 2021.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019.

C.-Y. Wei, Y.-T. Hong, and C.-J. Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

C.-Y. Wei, M. Jafarnia-Jahromi, H. Luo, H. Sharma, and R. Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.

C.-Y. Wei, M. Jafarnia-Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Q. Xie, Y. Chen, Z. Wang, and Z. Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pages 3674–3682. PMLR, 2020.

Y. Yang and J. Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, 2019.

K. Zhang, S. M. Kakade, T. Başar, and L. F. Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

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

Z. Zhang and X. Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, 2019.

Y. Zhou, J. Li, and J. Zhu. Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information. In *International Conference on Learning Representations*, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]