# Efficient Posterior Sampling in Competitive Reinforcement Learning: Function Approximation and Partial Observability

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 7, 2024

**Abstract**

Multi-agent reinforcement learning (MARL) has seen remarkable advances in various domains, such as autonomous driving and competitive games, where agents interact in shared environments with intertwined objectives. This paper investigates the problem of competitive reinforcement learning (RL) under function approximation and partial observation. Specifically, we propose a posterior sampling framework for two-player zero-sum Markov games (MGs), where agents operate with potentially incomplete observations of the state space. Our approach leverages two novel complexity measures—Self-Play Generalized Eluder Coefficient (GEC) and Adversarial GEC—that capture the exploration-exploitation tradeoff in competitive settings with function approximation. We design sample-efficient algorithms for both self-play and adversarial setups, achieving sublinear regret bounds. These results hold under general function approximation, covering important cases such as linear mixture MGs and partially observable MGs (POMGs). To the best of our knowledge, this is the first model-based posterior sampling framework addressing partial observability in competitive RL with provable sample efficiency.

**Keywords:** Posterior Sampling, Markov Games, Reinforcement Learning, Function Approximation, Partial Observability, Competitive RL

## 1 Introduction

Multi-agent reinforcement learning (MARL) has become an essential framework for addressing sequential decision-making problems where multiple agents interact within a shared environment, with each agent's actions affecting the others in a coupled manner. In a competitive reinforcement learning (RL) setting, the goal of each player is to maximize their own cumulative gains (or rewards), while often minimizing the cumulative gains (or maximizing the losses) of their opponents. Recent practical successes in domains such as autonomous driving, Go, StarCraft, and Poker have showcased the potential of MARL in solving complex, real-world problems.

The theoretical understanding of MARL, particularly for competitive environments modeled as Markov games (MGs), has lagged behind empirical breakthroughs. Much of the existing literature focuses on fully observable states and relies on optimism-based exploration techniques, which require careful design of bonus functions and optimization. However, many practical applications involve partial observability, where agents can only access limited information about the underlying states of the system. Moreover, there is a growing interest in understanding competitive RL under general function approximation, a key challenge when the state-action space is large or continuous.

In this paper, we address these gaps by proposing a novel posterior sampling framework for competitive RL with both full and partial observability under function approximation. We aim to develop sample-efficient algorithms for two-player zero-sum MGs, focusing on self-play and adversarial learning setups. Posterior sampling offers an alternative to optimism-based exploration methods, avoiding the need for model-specific bonuses and simplifying the computational complexity of exploration in RL. Our approach introduces two new complexity measures, the Self-Play Generalized Eluder Coefficient (GEC) and the Adversarial GEC, which generalize existing concepts from single-agent RL to competitive, multi-agent scenarios.

**Backgrounds.** Recent years have been tremendous practical successes of MARL in a variety of application domains, such as autonomous driving [SSSS16], Go [SHM$^+$16], StarCraft [VEB$^+$17], Dota2 [BBC$^+$19] and Poker [BS19]. These successes are attributed to advanced MARL algorithms that can coordinate multiple players by exploiting potentially partial observations of the latent states and employ powerful function approximators (neural networks in particular), which empower us to tackle practical problems with large state spaces.

Apart from the empirical success, there is a growing body of literature on establishing theoretical guarantees for Markov games (MGs) [Sha53] – a standard framework for describing the dynamics of competitive RL. In particular, [XCWY20, CZG22, JLY21, HLWY21, ZLY22] extend the works in single-agent reinforcement learning (RL) with function approximation [JKA$^+$17, SJK$^+$19, WSY20a, JYWJ20, AJS$^+$20, CYJW20, DMZZ21, JLM21, DKL$^+$21] by developing sample-efficient algorithms that are capable to solve two-player zero-sum MGs with function approximation. In addition, as opposed to the aforementioned literature on MGs assuming the state of players is fully observable, the recent work [LSJ22] analyze Markov games under partial observability [LSJ22], i.e., the complete information about underlying states is lacking. However, most of the existing works are built upon the principle of "optimism in the face of uncertainty" (OFU) [LS20] for exploration. Furthermore, from a practical perspective, achieving optimism often requires explicit construction of bonus functions, which are often designed in a model-specific fashion and computationally challenging to implement.

Another promising strand of exploration techniques is based on posterior sampling, which is shown by previous works on bandits [CL11] and RL [OBPVR16] to perform better than OFU-based algorithms. Meanwhile, posterior sampling methods, unlike OFU-based algorithms [JLM21, DKL$^+$21] that need to solve complex optimization problems to achieve optimism, can be efficiently implemented by ensemble approximations [OBPVR16, LVR17, CCML18, NKLK20] and stochastic gradient Langevin dynamics (SGLD) [WT11]. Despite the superiority of posterior sampling, its theoretical understanding in MARL remains limited. The only exception is [XZS$^+$22b], which proposes a model-free posterior sampling algorithm for zero-sum MGs with general function approximation. However, [XZS$^+$22b] cannot capture some common tractable competitive RL models with a model-based nature, such as linear mixture MGs [CZG22] and low witness rank MGs [HLWY21]. Moreover, their result is restricted to the fully observable MGs without handling the partial observability of the players' states. Therefore, we raise the following question: ***Can we design provably sample-efficient posterior sampling algorithms for competitive RL with even partial observations under general function approximation?***

Concretely, the above question poses three major challenges. First, despite the success of the OFU principle in partially observable Markov games (POMGs), it remains elusive how to incorporate the partial observations into the posterior sampling framework under a MARL setting with provably efficient exploration. Second, it is also unclear whether there is a generic function approximation condition that can cover more known classes in both full and partial observable MARL and is meanwhile compatible with the posterior sampling framework. Third, with the partial observation and function approximation, it is challenging to explore how we can solve MGs under the setups of self-play, where all players can be coordinated together, and adversarial learning, where the opponents' policies are adversarial and uncontrollable by the learner. Our work takes an initial step towards tackling such challenges by concentrating on the typical competitive RL scenario, the two-player zero-sum MG, and proposing statistically efficient posterior sampling algorithms under function approximation that can solve both self-play and adversarial MGs with full and partial observations.

**Contributions.** Our contributions are summarized as follows:

- We introduce two novel complexity measures—Self-Play Generalized Eluder Coefficient (GEC) and Adversarial GEC—that capture the exploration-exploitation tradeoff in competitive reinforcement learning (RL) with function approximation. These measures apply to a wide range of models, including linear MGs, linear mixture MGs, weakly revealing POMGs, and decodable POMGs.
- We propose a model-based posterior sampling algorithm for self-play in competitive RL, which effectively handles both fully and partially observable setups through carefully designed likelihood functions.
- We extend the posterior sampling framework to adversarial learning settings, where one player's policies may be adversarial or unknown, and show that the algorithm achieves sublinear regret bounds.
- Our framework is highly general, supporting various function approximation settings, and provides the first model-based posterior sampling approach that is sample-efficient for learning MGs with function approximation, under both self-play and adversarial setups. We also provide sublinear regret bounds scaling with the number of episodes $T$, the GEC $d_{\mathrm{GEC}}$, and the coverage of the optimal model by the initial sampling distribution.

**Organization** The remainder of the paper is organized as follows. In Section 2, we introduce the formal setup for two-player zero-sum MGs, function approximation, and the new complexity conditions. Section 3 presents our model-based posterior sampling algorithm for the self-play setting, along with theoretical regret analysis. Section 4 extends the algorithm to the adversarial setting and provides corresponding regret bounds. We conclude with related works and potential future directions in Section 5.

**Related Works.** There is a large body of literature studying MGs, especially zero-sum MGs. In the self-play setting, many papers have focused on solving approximate NE in tabular zero-sum MGs [BJ20, BJY20, XCWY20, QWY$^+$21, LYBJ21], zero-sum MGs with linear function approximation [XCWY20, CZG22], zero-sum MGs with low-rank structures [QWB$^+$22, ZRY$^+$22, NSZ$^+$22], and zero-sum MGs with general function approximation [JLY21, HLWY21, XZS$^+$22b]. On the other hand, there are also several recent papers focusing on the adversarial setting [XCWY20, TWYS21, JLY21, HLWY21] aim to learn the Nash value under the setting of unrevealed opponent's policies, where the adversarial policies of the opponent are unobservable. In addition, another line of adversarial MGs concentrates on a revealed policy setting, where the opponent's full policy can be observed, leading to efficiently learning a sublinear regret comparing against the best policy in hindsight. Particularly, [LWJ22] and [ZLY22] develop efficient algorithms in tabular and function approximation settings, respectively. Our approach focuses on the unrevealed policy setting, which is considered to be a more practical setup. There are also works studying MGs from various aspects [SMB21, JLWY21, MB21, ZYWJ23, JMS22, DGZ22, QYWY21, BLZZ23, ZXT$^+$22, CD22, XZS$^+$22a, YZW$^+$23], such as multi-player general-sum MGs, reward-free MGs, MGs with delayed feedback, and offline MGs, which are beyond the scope of our work. Most of the aforementioned works follow the OFU principle and differ from our posterior sampling methods. The recent work [XZS$^+$22b] proposes a model-free posterior sampling algorithm for two-player zero-sum MGs but is limited to the self-play setting with fully observable states. Moreover, their work requires a strong Bellman completeness assumption that is restrictive compared to only requiring realizability in our work, mainly due to the monotonicity, meaning that adding a new function to the function class may violate it. Many model-based models like linear mixture MGs [CZG22] and low witness rank MGs [HLWY21] are not Bellman-complete, so they cannot be captured by [XZS$^+$22b]. Without the completeness assumption, our model-based posterior sampling approaches can solve a rich class of tractable MGs, including linear mixture MGs, low witness rank MGs [HLWY21], and even POMGs, tackling both self-play and adversarial learning settings.

Our work is related to a line of research on posterior sampling methods in RL. For single-agent RL, most existing works such as [RVR14] analyze the Bayesian regret bound. There are also some works [AJ17, Rus19, ZBB$^+$20] focusing on the frequentist (worst-case) regret bound. Our work is more closely related to the recently developed feel-good Thompson sampling technique proposed by [Zha22] for the frequentist regret bound, and its extension to single-agent RL [DMZZ21, AZ22a, AZ22b] and two-player zero-sum MGs [XZS$^+$22b].

Our work is also closely related to the line of research on function approximation in RL. Such a line of works proposes algorithms for efficient policy learning under diverse function approximation classes, spanning from linear Markov decision processes (MDPs) [JYWJ20], linear mixture MDPs [AJS$^+$20, ZGS21] to nonlinear and general function classes, including, for instance, generalized linear MDPs [WWDK19], kernel and neural function classes [YJW$^+$20], bounded eluder dimension [OVR14, WSY20b], Bellman rank [JKA$^+$17], witness rank [SJK$^+$19], bellman eluder dimension [JLM21], bilinear [DKL$^+$21], decision-estimation coefficient [FKQR21, CMB22], decoupling coefficient [DMZZ21], admissible Bellman characterization [CLY$^+$22], and GEC [ZXZ$^+$22] classes.

The research on partial observability in RL [GDB16] is closely related to our work. The works [KAL16, JKKL20] show that learning history-dependent policies generally can cause an exponential sample complexity. Thus, many recent works focus on analyzing tractable subclasses of partially observable Markov decision processes (POMDPs), which includes weakly revealing POMDPs [JKKL20, LCSJ22], observable POMDPs [GMR22b, GMR22a], decodable POMDPs [DKJ$^+$19, EJKM22], low-rank POMDPs[WCYW], regular PSR [ZUSL22], PO-bilinear class [USL$^+$22b], latent MDP with sufficient tests [KECM21], B-stable PSR [CBM22], well-conditioned PSR [LNSJ22], POMDPs with deterministic transition kernels [JKKL20, USL$^+$22a], and GEC [ZXZ$^+$22]. Nevertheless, in contrast to our work which focuses on the two-player competitive setting with partial observation, these papers merely consider the single-agent setting. The recent research [LSJ22] further generalizes weakly revealing POMDPs to its multi-agent counterpart, weakly revealing POMGs, in a general-sum multi-player setting based on the OFU principle. But when specialized to the two-player case, our work proposes a general function class that can subsume the class of weakly revealing POMGs as a special case. It would be intriguing to generalize our framework to the general-sum settings in the future.

**Notations.** We denote by $\mathrm{KL}(P||Q) = \mathbb{E}_{x\sim P}[\log(\mathrm{d}P(x)/\mathrm{d}Q(x))]$ the KL divergence and $D^2_{\mathrm{He}}(P,Q) = 1/2\cdot\mathbb{E}_{x\sim P}(\sqrt{\mathrm{d}Q(x)/\mathrm{d}P(x)}-1)^2$ the Hellinger distance. We denote by $\Delta_{\mathcal{X}}$ the set of all distributions over $\mathcal{X}$ and $\mathrm{Unif}(\mathcal{X})$ the uniform distribution over $\mathcal{X}$. We let $x \wedge y$ be $\min\{x, y\}$.

## 2 Problem Setup

We introduce the basic concept of the two-player zero-sum Markov game (MG), function approximation, and the new complexity conditions for function approximation. Concretely, we study two typical classes of MGs, i.e., fully observable MGs and partially observable MGs, as defined below.

**Fully Observable Markov Game.** We consider an episodic two-player zero-sum fully observable Markov game (FOMG[1]) specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r, H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ and $\mathcal{B}$ are the action spaces of Players 1 and 2 respectively, $H$ is the length of the episode. We denote by $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ the transition kernel with $\mathbb{P}_h(s'|s, a, b)$ specifying the probability (density) of transitioning from state $s$ to state $s'$ given Players 1 and 2's actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$ at step $h$. We denote the reward function as $r = \{r_h\}_{h=1}^H$ with $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto [0, 1]$ being the reward received by players at step $h$. We define $\pi = \{\pi_h\}_{h=1}^H$ and $\nu = \{\nu_h\}_{h=1}^H$ as *Markovian* policies for Players 1 and 2, i.e., $\pi_h(a|s)$ and $\nu_h(b|s)$ are the probability of taking action $a$ and $b$ conditioned on the current state $s$ at step $h$. Without loss of generality, we assume the initial state $s_1$ is fixed for each episode. We consider a realistic setting where the transition kernel $\mathbb{P}$ is *unknown* and thereby needs to be approximated using the collected data.

**Partially Observable Markov Game.** This paper further studies an episodic zero-sum partially observable Markov game (POMG), which is distinct from the FOMG setup in that the state $s$ is not directly observable. In particular, a POMG is represented by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{O}, \mathbb{P}, \mathbb{O}, \mu, r, H)$, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{B}$ , $H$, and $\mathbb{P}$ are similarly the state and action spaces, the episode length, and the transition kernel. Here $\mu_1(\cdot)$ denotes the initial state distribution. We denote by $\mathbb{O} := \{\mathbb{O}_h\}_{h=1}^H$ the emission kernel so that $\mathbb{O}_h(o|s)$ is the probability of having a partial observation $o \in \mathcal{O}$ at state $s$ with $\mathcal{O}$ being the observation space. Since we only have an observation $o$ of a state, the reward function is defined as $r := \{r_h\}_{h=1}^H$ with $r_h(o, a, b) \in [0, 1]$ depending on actions $a, b$ and the observation $o$, and the policies for players are defined as $\pi = \{\pi_h\}_{h=1}^H$ and $\nu = \{\nu_h\}_{h=1}^H$, where $\pi_h(a_h|\tau_{h-1}, o_h)$ and $\nu_h(b_h|\tau_{h-1}, o_h)$ is viewed as the probability of taking actions $a_h$ and $b_h$ depending on all histories $(\tau_{h-1}, o_h)$. Here we let $\tau_h := (o_1, a_1, b_1 \ldots, o_h, a_h, b_h)$. Then, in contrast to FOMGs, the policies in POMGs are *history-dependent*, defined on all prior observations and actions rather than the current state $s$. We define $\mathbf{P}_h^{\pi,\nu}(\tau_h) := \int_{\mathcal{S}^h} \mu_1(s_1) \prod_{h'=1}^{h-1}[\mathbb{O}_{h'}(o_{h'}|s_{h'})\pi_{h'}(b_{h'}|\tau_{h'-1}, o_{h'})\nu_{h'}(b_{h'}|\tau_{h'-1}, o_{h'})\mathbb{P}_{h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'})]\mathbb{O}_h(o_h|s_h)\mathrm{d}s_{1:h}$, which is the joint distribution of $\tau_h$ under the policy pair $(\pi, \nu)$. Removing policies in $\mathbf{P}_h^{\pi,\nu}$, we define the function $\mathbf{P}_h(\tau_h) := \int_{\mathcal{S}^h} \mu_1(s_1) \prod_{h'=1}^{h-1}[\mathbb{O}_{h'}(o_{h'}|s_{h'})\mathbb{P}_{h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'})]\mathbb{O}_h(o_h|s_h)\mathrm{d}s_{1:h}$. We assume that the parameters $\theta := (\mu_1, \mathbb{P}, \mathbb{O})$ are *unknown* and thus $\mathbf{P}_h$ is *unknown* as well, which should be approximated in algorithms via online interactions.

**Online Interaction with the Environment.** In POMGs, at step $h$ of episode $t$ of the interaction, players take actions $a_h^t \sim \pi_h^t(\cdot|\tau_{h-1}^t, o_h^t)$ and $b_h^t \sim \nu_h^t(\cdot|\tau_{h-1}^t, b_{h-1}^t, o_h^t)$ depending on their action and observation histories, receiving a reward $r_h(o_h^t, a_h^t, b_h^t)$ and transitions from the latent

[1] FOMG is typically referred to as MG by most literature. We adopt FOMG to differentiate it from POMG.

state $s_h^t$ to $s_{h+1}^t \sim \mathbb{P}_h(\cdot \mid s_h^t, a_h^t, b_h^t)$ with an observation $o_{h+1}^t \sim \mathbb{O}_h(\cdot|s_h^t)$ generated. When the underlying state $s_h^t$ is observable and the policies become Markovian, we have actions $a_h^t \sim \pi_h^t(s_h^t)$ and $b_h^t \sim \nu_h^t(s_h^t)$ and the reward $r_h(s_h^t, a_h^t, b_h^t)$. Then, it reduces to the interaction process under the FOMG setting.

**Value Function, Best Response, and Nash Equilibrium.** To characterize the learning objective and the performance of the algorithms, we define the value function as the expected cumulative rewards under the policy pair $(\pi, \nu)$ starting from the initial step $h = 1$. For FOMGs, we define the value function as $V^{\pi,\nu} := \mathbb{E}[\sum_{h=1}^H r_h(s_h, a_h, b_h) \mid s_1, \pi, \nu, \mathbb{P}]$, where the expectation is taken over all the randomness induced by $\pi$, $\nu$, and $\mathbb{P}$. For POMG, we define the value function as $V^{\pi,\nu} := \mathbb{E}[\sum_{h=1}^H r_h(o_h, a_h, b_h) \,|\pi, \nu, \theta]$, with the expectation taken for $\pi$, $\nu$, and $\theta$.

Our work studies the competitive setting of RL, where Player 1 (*max-player*) aims to maximize the value function $V^{\pi,\nu}$ while Player 2 (*min-player*) aims to minimize it. With the defined value function, given a policy pair $(\pi, \nu)$, we define their *best responses* respectively as $\mathtt{br}(\pi) \in \arg\min_\nu V^{\pi,\nu}$ and $\mathtt{br}(\nu) \in \arg\max_\pi V^{\pi,\nu}$. Then, we say a policy pair $(\pi^*, \nu^*)$ is a *Nash equilibrium* (NE) if

$$V^{\pi^*,\nu^*} = \max_\pi \min_\nu V^{\pi,\nu} = \min_\nu \max_\pi V^{\pi,\nu}$$

Thus, it always holds that $\pi^* = \mathtt{br}(\nu^*)$ and $\nu^* = \mathtt{br}(\pi^*)$. For abbreviation, we denote $V^* = V^{\pi^*,\nu^*}$, $V^{\pi,*} = \min_\nu V^{\pi,\nu}$, and $V^{*,\nu} = \max_\pi V^{\pi,\nu}$, which implies $V^* = V^{\pi^*,*} = V^{*,\nu^*}$ for NE $(\pi^*, \nu^*)$. Moreover, we define the policy pair $(\pi, \nu)$ as an $\varepsilon$-approximate NE if it satisfies $V^{*,\nu} - V^{\pi,*} \leq \varepsilon$.

**Function Approximation.** Since the environment is unknown to players, the model-based RL setting requires us to learn the true model of the environment, $f^*$, via (general) function approximation. We use the functions $f$ lying in a general model function class $\mathcal{F}$ to approximate the environment. We make a standard realizability assumption on the relationship between the model class and the true model.

**Assumption 1** (Realizability)**.** *For a model class $\mathcal{F}$, the true model $f^*$ satisfies $f^* \in \mathcal{F}$.*

In our work, the true model $f^*$ represents the transition kernel $\mathbb{P}$ for the FOMG and $\theta$ for the POMG. For any $f \in \mathcal{F}$, we let $\mathbb{P}_f$ and $\theta_f = (\mu_f, \mathbb{P}_f, \mathbb{O}_f)$ be the models under the approximation function $f$ and $V_f^{\pi,\nu}$ the value function associated with $f$. For POMGs, we denote $\mathbf{P}_{f,h}^{\pi,\nu}$ and $\mathbf{P}_{f,h}$ as $\mathbf{P}_h^{\pi,\nu}$ and $\mathbf{P}_h$ under the model $f$.

**MGs with Self-Play and Adversarial Learning.** Our work investigates two important MG setups for competitive RL, which are the self-play setting and the adversarial setting. In the self-play setting, the learner can control *both* players together to execute the proposed algorithms to learn an approximate NE. Therefore, our objective is to design sample-efficient algorithms to generate a sequence of policy pairs $\{(\pi^t, \nu^t)\}_{t=1}^T$ in $T$ episodes such that the following regret can be minimized,

$$\mathrm{Reg}^{\mathrm{sp}}(T) := \textstyle\sum_{t=1}^T \left[V_{f^*}^{*,\nu^t} - V_{f^*}^{\pi^t,*}\right]$$

In the adversarial setting, we can no longer coordinate both players, and only *single* player is controllable. Under such a circumstance, the opponent plays arbitrary and even adversarial policies.

Wlog, suppose that the main player is the max-player with the policies $\{\pi^t\}_{t=1}^T$ generated by a carefully designed algorithm and the opponent is min-player with arbitrary policies $\{\nu^t\}_{t=1}^T$. The objective of the algorithm is to learn policies $\{\pi^t\}_{t=1}^T$ to maximize the overall cumulative rewards in the presence of an adversary. To measure the performance of algorithms, we define the following regret for the adversarial setting by comparing the learned value against the Nash value, i.e.,

$$\mathrm{Reg}^{\mathrm{adv}}(T) = \textstyle\sum_{t=1}^T \left[V_{f^*}^* - V_{f^*}^{\pi^t,\nu^t}\right]$$

# 3 Model-Based Posterior Sampling for the Self-Play Setting

We propose algorithms aiming to generate a sequence of policy pairs $\{(\pi^t, \nu^t)\}_{t=1}^T$ by controlling the learning process of both players such that the regret $\mathrm{Reg}^{\mathrm{sp}}(T)$ is small. Such a regret can be decomposed into two parts, namely $\sum_{t=1}^T [V_{f^*}^* - V_{f^*}^{\pi^t,*}]$ and $\sum_{t=1}^T [V_{f^*}^{*,\nu^t} - V_{f^*}^*]$, which inspires our to design algorithms for learning $\{\pi^t\}_{t=1}^T$ and $\{\nu^t\}_{t=1}^T$ separately by targeting at minimizing these two parts respectively. Due to the symmetric structure of such a game learning problem, we propose the algorithm to learn $\{\pi^t\}_{t=1}^T$ as summarized in Algorithm 1. The algorithm for learning $\{\nu^t\}_{t=1}^T$ can be proposed in a symmetric way in Algorithm 3, which is deferred to Appendix A. Our proposed algorithm features an integration of the model-based posterior sampling and the exploiter-guided self-play in a multi-agent learning scenario. In Algorithm 1, Player 1 is the main player, while Player 2 is called the exploiter, who assists the learning of the main player by exploiting her weakness.

**Posterior Sampling for the Main Player.** The posterior sampling constructs a posterior distribution $p^t(\cdot|Z^{t-1})$ over the function class $\mathcal{F}$ each round based on collected data and a pre-specified prior distribution $p^0(\cdot)$, where $Z^{t-1}$ denotes the random history up to the end of the $(t-1)$-th episode. For ease of notation, hereafter, we omit $Z^{t-1}$ in the posterior distribution. Most recent literature shows that adding an optimism term in the posterior distribution can lead to sample-efficient RL algorithms. Thereby, we define the distribution $p^t(\cdot)$ over the function class $\mathcal{F}$ for the main player as in Line 3 of Algorithm 1, which is proportional to $p^0(f) \exp[\gamma V_f^* + \sum_{\tau=1}^{t-1} \sum_{h=1}^H L_h^\tau(f)]$. Here, $\gamma V_f^*$ serves as the optimism term, and $L_h^\tau(f)$ is the likelihood function built upon the pre-collected data. Such a construction of $p^t(\cdot)$ indicates that we will assign a higher probability (density) to a function $f$, which results in higher values of the combination of the optimism term and the likelihood function. We sample a model $\overline{f}^t$ from the distribution $p^t(\cdot)$ over the model class and learn the policy $\pi^t$ for the main player such that $(\pi^t, \overline{\nu}^t)$ is the NE of the value function under the model $\overline{f}^t$ in Line 4, where $\overline{\nu}^t$ is a dummy policy and only used in our theoretical analysis.

**Posterior Sampling for the Exploiter.** The exploiter aims to track the best response of $\pi^t$ to assist learning a low regret. The best response of $\pi^t$ generated by the exploiter is nevertheless based on a value function under a different model than $\overline{f}^t$. Specifically, for the exploiter, we define the posterior sampling distribution $q^t(\cdot)$ using an optimism term $-\gamma V_f^{\pi^t,*}$ and the summation of likelihood functions, i.e., $\sum_{\tau=1}^{t-1} \sum_{h=1}^H L_h^\tau(f)$, along with a prior distribution $q^0(\cdot)$, in Line 5 of Algorithm 1. The negative term $-\gamma V_f^{\pi^t,*}$ favors a model with a low value and is thus optimistic from the exploiter's perspective but pessimistic for the main player. We then sample a model $\underline{f}^t$ from $q^t(\cdot)$ and compute the best response of $\pi^t$, denoted as $\underline{\nu}^t$, under the model $\underline{f}^t$ as in Line 7.

**Algorithm 1** Model-Based Posterior Sampling for Self-Play (Max-Player)

```
1: Input: Model class $\mathcal{F}$, prior distributions $p^0$ and $q^0$, $\gamma_1$, and $\gamma_2$.
2: for $t = 1, \ldots, T$ do
3:     Draw a model $\overline{f}^t \sim p^t(\cdot)$ with defining $p^t(f) \propto p^0(f)\exp[\gamma_1 V_f^* + \sum_{\tau=1}^{t-1}\sum_{h=1}^{H} L_h^\tau(f)]$
4:     Compute $\pi^t$ by letting $(\pi^t, \overline{\nu}^t)$ be the NE of $V_{\overline{f}^t}^{\pi,\nu}$
5:     Draw a model $\underline{f}^t \sim q^t(\cdot)$ with defining $q^t(f) \propto q^0(f)\exp[-\gamma_2 V_f^{\pi^t,*} + \sum_{\tau=1}^{t-1}\sum_{h=1}^{H} L_h^\tau(f)]$
6:     Compute $\underline{\nu}^t$ by letting $\underline{\nu}^t$ be the best response of $\pi^t$ w.r.t. $V_{\underline{f}^t}^{\pi,\nu}$
7:     Collect data $\mathcal{D}^t$ by executing the joint exploration policy $\sigma^t$
8:     Define the likelihood functions $\{L_h^t(f)\}_{h=1}^H$ using the collected data $\mathcal{D}^t$
9: end for
10: Return: $(\pi^1, \ldots, \pi^T)$
```

**Data Sampling and Likelihood Function.** With the learned $\pi^t$ and $\underline{\nu}^t$, we define a joint exploration policy $\sigma^t$ in Line 7 of Algorithm 1, by executing which we can collect a dataset $\mathcal{D}^t$. We are able to further construct the likelihood functions $\{L_h^t(f)\}_{h=1}^H$ in Line 8 using $\mathcal{D}^t$. Different game settings require specifying diverse exploration policies $\sigma^t$ and likelihood functions $\{L_h^t(f)\}_{h=1}^H$. Particularly, for the game classes mainly discussed in this work, we set $\sigma^t = (\pi^t, \underline{\nu}^t)$ for both FOMGs and POMGs. In FOMGs, we let $\mathcal{D}^t = \{(s_h^t, a_h^t, b_h^t, s_{h+1}^t)\}_{h=1}^H$, where for each $h \in [H]$, the data point $(s_h^t, a_h^t, b_h^t, s_{h+1}^t)$ is collected by executing $\sigma^t$ to the $h$-th step of the game. The corresponding likelihood function is defined using the transition kernel as

$$L_h^t(f) = \eta \log \mathbb{P}_{f,h}(s_{h+1}^t \,|\, s_h^t, a_h^t, b_h^t) \tag{1}$$

Furthermore, under the POMG setting, we let the dataset be $\mathcal{D}^t = \{\tau_h^t\}_{h=1}^H$, where the data point $\tau_h^t = (o_1^t, a_1^t, b_1^t \ldots, o_h^t, a_h^t, b_h^t)$ is collected by executing $\sigma^t$ to the $h$-th step of the game for each $h \in [H]$. We further define the associated likelihood function as

$$L_h^t(f) = \eta \log \mathbf{P}_{f,h}(\tau_h^t) \tag{2}$$

Such a construction of the likelihood function in a log-likelihood form can result in learning a model $f$ well approximating the true model $f^*$ measured via the Hellinger distance.

## 3.1 Regret Analysis for the Self-play Setting

Our regret analysis is based on a novel structural complexity condition for multi-agent RL and a quantity to measure how the well the prior distributions cover the optimal model $f^*$. We first define the following condition for the self-play setting.

**Definition 1** (Self-Play GEC)**.** *For any sequences of functions $f^t, g^t \in \mathcal{F}$, suppose that a pair of policies $(\pi^t, \nu^t)$ satisfies: **(a)** $\pi^t = \mathrm{argmax}_\pi \min_\nu V_{f^t}^{\pi,\nu}$ and $\nu^t = \mathrm{argmin}_\nu V_{g^t}^{\pi^t,\nu}$, or **(b)** $\nu^t = \mathrm{argmin}_\nu \max_\pi V_{f^t}^{\pi,\nu}$ and $\pi^t = \mathrm{argmax}_\pi V_{g^t}^{\pi,\nu^t}$. Denoting the joint exploration policy as $\sigma^t$ depending on $f^t$ and $g^t$, for any $\rho \in \{f, g\}$ and $(\pi^t, \nu^t)$ following **(a)** and **(b)**, the self-play GEC $d_{\mathrm{GEC}}$ is defined as the minimal constant $d$ satisfying*

$$\left|\sum_{t=1}^T \left(V_{\rho^t}^{\pi^t,\nu^t} - V_{f^*}^{\pi^t,\nu^t}\right)\right| \le \left[d\sum_{h=1}^H \sum_{t=1}^T \left(\sum_{\tau=1}^{t-1} \mathbb{E}_{(\sigma^\tau,h)}\ell(\rho^t, \xi_h^\tau)\right)\right]^{\frac{1}{2}} + 2H(dHT)^{\frac{1}{2}} + \epsilon HT$$

Our definition of self-play GEC is inspired by [ZXZ+22] for the single-agent RL. Then, it shares an analogous meaning to the single-agent GEC. Here $(\sigma^\tau, h)$ implies running the joint exploration policy $\sigma^\tau$ to step $h$ to collect a data point $\xi_h^\tau$. The LHS of the inequality is viewed as the prediction error and the RHS is the training error defined on a loss function $\ell$ plus a burn-in error $2H(dHT)^{\frac{1}{2}} + \epsilon HT$ that is non-dominating when $\epsilon$ is small. The loss function $\ell$ and $\epsilon$ can be problem-specific. We determine $\ell(f, \xi_h)$ for FOMGs with $\xi_h = (s_h, a_h)$ and POMGs with $\xi_h = \tau_h$ respectively as

$$\text{FOMG: } D^2_{\mathrm{He}}(\mathbb{P}_{f,h}(\cdot|\xi_h), \mathbb{P}_{f^*,h}(\cdot|\xi_h)), \quad \text{POMG: } 1/2 \cdot \left(\sqrt{\mathbf{P}_{f,h}(\xi_h)/\mathbf{P}_{f^*,h}(\xi_h)} - 1\right)^2 \tag{3}$$

such that $\mathbb{E}_{(\sigma,h)}[\ell(f, \xi_h)] = D^2_{\mathrm{He}}(\mathbf{P}^\sigma_{f^*,h}, \mathbf{P}^\sigma_{f,h})$ for POMGs. The intuition for GEC is that if hypotheses have a small training error on a well-explored dataset, then the out-of-sample prediction error is also small, which characterizes the hardness of environment exploration.

Since the posterior sampling steps in our algorithms depend on the initial distributions $p^0$ and $q^0$, we define the following quantity to measure how well the prior distributions $p^0$ and $q^0$ cover the optimal model $f^* \in \mathcal{F}$, which is also a multi-agent generalization of its single-agent version [AZ22a, ZXZ+22].

**Definition 2** (Prior around True Model)**.** *Given $\beta > 0$ and any distribution $p^0 \in \Delta_{\mathcal{F}}$, we define*

$$\omega(\beta, p^0) = \inf_{\varepsilon > 0} \{\beta\varepsilon - \log p^0[\mathcal{F}(\varepsilon)]\}$$

*where $\mathcal{F}(\varepsilon) := \{f \in \mathcal{F} \ : \ \sup_{h,s,a,b} \mathrm{KL}^{\frac{1}{2}}(\mathbb{P}_{f^*,h}(\cdot \mid s,a,b) \| \mathbb{P}_{f,h}(\cdot \mid s,a,b)) \le \varepsilon\}$ for FOMGs and $\mathcal{F}(\varepsilon) := \{f \in \mathcal{F} \ : \ \sup_{\pi,\nu} \mathrm{KL}^{\frac{1}{2}}(\mathbf{P}^{\pi,\nu}_{f^*,H} \| \mathbf{P}^{\pi,\nu}_{f,H}) \le \varepsilon\}$ for POMGs.*

When the model class $\mathcal{F}$ is a finite space, if let $p^0 = \mathrm{Unif}(\mathcal{F})$, we simply know that $\omega(\beta, p^0) \le \log|\mathcal{F}|$ where $|\mathcal{F}|$ is the cardinality of $\mathcal{F}$. Furthermore, for an infinite function class $\mathcal{F}$, the term $\log|\mathcal{F}|$ can be substituted by a quantity having logarithmic dependence on the covering number of the function class $\mathcal{F}$. With the multi-agent GEC condition and the definition of $\omega$, we have the following regret bound for both FOMGs and POMGs.

**Proposition 1.** *Letting $\eta = 1/2$, $\gamma_1 = 2\sqrt{\omega(4HT, p^0)T/d_{\mathrm{GEC}}}$, $\gamma_2 = 2\sqrt{\omega(4HT, q^0)T/d_{\mathrm{GEC}}}$, $\epsilon = 1/\sqrt{HT}$ in Definition 1, when $T \ge \max\{4H^2\omega(4HT, p^0)/d_{\mathrm{GEC}}, 4H^2\omega(4HT, q^0)/d_{\mathrm{GEC}}, d_{\mathrm{GEC}}/H\}$, under both FOMG and POMG settings, Algorithm 1 admits the following regret bound,*

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{sp}}_1(T)] := \mathbb{E}[\textstyle\sum_{t=1}^T (V^*_{f^*} - V^{\pi^t,*}_{f^*})] \le 6\sqrt{d_{\mathrm{GEC}} HT \cdot [\omega(4HT, p^0) + \omega(4HT, q^0)]}$$

This proposition gives the upper bound $\mathbb{E}[\mathrm{Reg}^{\mathrm{sp}}_1(T)]$ following the updating rules in Algorithm 1 when the max-player is the main player. As Algorithm 3 is symmetric to Algorithm 1, we obtain the following regret bound of $\mathbb{E}[\mathrm{Reg}^{\mathrm{sp}}_2(T)]$ for Algorithm 3 when the min-player is the main player.

**Proposition 2.** *Under the same parameter settings as Proposition 1, Algorithm 3 admits the following regret bound,*

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{sp}}_2(T)] := \mathbb{E}[\textstyle\sum_{t=1}^T (V^{*,\nu^t}_{f^*} - V^*_{f^*})] \le 6\sqrt{d_{\mathrm{GEC}} HT \cdot [\omega(4HT, p^0) + \omega(4HT, q^0)]}$$

Combining the results of Propositions 1 and 2, due to $\mathrm{Reg}^{\mathrm{sp}}(T) = \mathrm{Reg}^{\mathrm{sp}}_1(T) + \mathrm{Reg}^{\mathrm{sp}}_2(T)$, we obtain the following overall regret when running Algorithms 1 and 3 together.

**Theorem 3.1.** *Under the settings of Propositions 1 and 2, executing both Algorithms 1 and 3 leads to*

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{sp}}(T)] \leq 12\sqrt{d_{\mathrm{GEC}} HT \cdot [\omega(4HT, p^0) + \omega(4HT, q^0)]}$$

The above results indicate that the proposed posterior sampling self-play algorithms (Algorithms 1 and 3) separately admit a sublinear dependence on GEC $d_{\mathrm{GEC}}$, the number of learning episodes $T$, as well as $\omega(4HT, p^0)$ and $\omega(4HT, q^0)$ for both FOMG and POMG settings. They lead to the same overall regret bound combining Propositions 1 and 2. In particular, when $\mathcal{F}$ is finite with $p^0 = q^0 = \mathrm{Unif}(\mathcal{F})$, Algorithms 1 and 3 admit regrets of $O(\sqrt{d_{\mathrm{GEC}} HT \cdot \log|\mathcal{F}|})$. The quantity $\omega$ can be associated with the log-covering number if $\mathcal{F}$ is infinite. Please see Appendix C for analysis.

# 4 Posterior Sampling for the Adversarial Setting

Without loss of generality, we assume that the max-player is the main agent and the min-player is the opponent. Under this setting, the goal of the main player is to maximize her cumulative rewards as much as possible, comparing against the value under the NE, i.e., $V_{f^*}^*$. We develop a novel algorithm for this setting as summarized in Algorithm 2. In our algorithm, the opponent's policy is assumed to be arbitrary and is also *not revealed* to the main player. The only information about the opponent is the current state or the partial observation of her state as well as the actions taken.

We adopt the optimistic posterior sampling approach for the main player with defining an optimism term as $\gamma V_f^*$ motivated by the above learning target, and the likelihood function $L_h^t(f)$ with $L_h^t(f) := \eta \log \mathbb{P}_{f,h}(s_{h+1}^t \,|\, s_h^t, a_h^t, b_h^t)$ in (1) for FOMGs and $L_h^t(f) = \eta \log \mathbf{P}_{f,h}(\tau_h^t)$ in (2) for POMGs respectively. The policy $\pi^t$ learned by the main player is from computing the NE of the value function under the current model $f^t$ sampled from the posterior distribution $p^t$. In addition, the joint exploration policy is set to be $\sigma^t = (\pi^t, \nu^t)$ where $\nu^t$ is the potentially adversarial policy played by the opponent. Thus, we can collect the data defined as $\mathcal{D}^t = \{(s_h^t, a_h^t, b_h^t, s_{h+1}^t)\}_{h=1}^H$ and $\mathcal{D}^t = \{\tau_h^t\}_{h=1}^H$ with $\tau_h^t = (o_1^t, a_1^t, b_1^t \ldots, o_h^t, a_h^t, b_h^t)$ for FOMGs and POMGs respectively, collected by executing $\sigma^t$ to the $h$-th step of the game for each $h \in [H]$.

**Remark.** In Algorithm 2, we define the joint exploration policy $\sigma^t = (\pi^t, \nu^t)$, which is the key to the success of the algorithm design under the adversarial setting, especially for POMGs. Under the single-agent setting, the prior work [ZXZ$^+$22] sets the exploration policy for a range of partially observable models subsumed by the PSR model as $\pi_{1:h-1}^t \circ_h \mathrm{Unif}(\mathcal{A})$, i.e., running $\pi^t$ for steps 1 to $h-1$ and then sampling the data at step $h$ by enforcing a uniform policy. Such an exploration scheme fails to work when facing an uncontrollable opponent who does not play a uniform policy at step $h$. Theoretically, we prove that employing policies $(\pi_{1:h}^t, \nu_{1:h}^t)$ for exploration without the uniform policy, the self-play and adversarial GEC conditions in Definitions 1 and 3 are still satisfied for a class of POMGs including weakly revealing and decodable POMGs. This eventually leads to a unified adversarial learning algorithm for both FOMGs and POMGs.

## 4.1 Regret Analysis for the Adversarial Setting

Before demonstrating our regret analysis, we first define a multi-agent GEC fitting the adversarial learning scenario. Considering that the opponent's policy is uncontrollable during the learning, we let $\{\nu^t\}_{t=1}^T$ be arbitrary, which is clearly distinguished from self-play GEC defined in Definition 1.

**Algorithm 2** Model-Based Posterior Sampling with Adversarial Opponent

```
Input: Model class $\mathcal{F}$, prior distributions $p^0$, and $\gamma$.
for $t = 1, \ldots, T$ do
    Draw a model $f^t \sim p^t(\cdot)$ with defining $p^t(f) \propto p^0(f)\exp[\gamma V_f^* + \sum_{\tau=1}^{t-1}\sum_{h=1}^{H} L_h^\tau(f)]$
    Compute $\pi^t$ by letting $(\pi^t, \overline{\nu}^t)$ be NE of $V_{f^t}^{\pi,\nu}$
    Opponent picks an arbitrary policy $\nu^t$
    Collect a trajectory $\mathcal{D}^t$ by executing the joint exploration policy $\sigma^t$
    Define the likelihood functions $\{L_h^t(f)\}_{h=1}^H$ using the collected data $\mathcal{D}^t$
end for
Return: $(\pi^1, \ldots, \pi^T)$
```

**Definition 3** (Adversarial GEC)**.** *For any sequence of functions $\{f^t\}_{t=1}^T$ with $f^t \in \mathcal{F}$ and any sequence of the opponent's policies $\{\nu^t\}_{t=1}^T$, suppose that the main player's policies $\{\mu^t\}_{t=1}^T$ are generated via $\mu^t = \operatorname{argmax}_\pi \min_\nu V_{f^t}^{\pi,\nu}$. Denoting the joint exploration policy as $\{\sigma^t\}_{t=1}^T$ depending on $\{f^t\}_{t=1}^T$, the adversarial GEC $d_{\mathrm{GEC}}$ is defined as the minimal constant $d$ satisfying*

$$\sum_{t=1}^T \left(V_{f^t}^{\pi^t,\nu^t} - V_{f^*}^{\pi^t,\nu^t}\right) \le \left[d\sum_{h=1}^H\sum_{t=1}^T\left(\sum_{\tau=1}^{t-1}\mathbb{E}_{(\sigma^\tau,h)}\ell(f^t,\xi_h^\tau)\right)\right]^{\frac{1}{2}} + 2H(dHT)^{\frac{1}{2}} + \epsilon HT$$

Our regret analysis for Algorithm 2 also depends on the quantity $\omega(\beta, p^0)$ that characterizes the coverage of the prior distribution $p^0$ on the true model $f^*$. Then, we have the following regret bound.

**Theorem 4.1.** *Letting $\eta = \frac{1}{2}$, $\gamma = 2\sqrt{\omega(4HT, p^0)T/d_{\mathrm{GEC}}}$, $\epsilon = 1/\sqrt{HT}$ in Definition 3, when $T \ge \max\{4H^2\omega(4HT,p^0)/d_{\mathrm{GEC}}, d_{\mathrm{GEC}}/H\}$, under both FOMG and POMG settings, Algorithm 2 admits the following regret bound,*

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{adv}}(T)] \le 4\sqrt{d_{\mathrm{GEC}}HT \cdot \omega(4HT, p^0)}$$

The above result indicates that we can achieve a meaningful regret bound by a posterior sampling algorithm with general function approximation, even when the opponent's policy is adversarial and her full policies $\nu^t$ are not revealed. This regret has a sublinear dependence on $d_{\mathrm{GEC}}$, the number of episodes $T$, as well as $\omega(4HT, p^0)$. Similarly, when $\mathcal{F}$ is finite, Algorithm 2 admits a regret of $O(\sqrt{d_{\mathrm{GEC}}HT \cdot \log|\mathcal{F}|})$. The term $\log|\mathcal{F}|$ can be the log-covering number of $\mathcal{F}$ if it is infinite.

## 5 Theoretical Analysis

This section presents several examples of tractable MG classes captured by the self-play and adversarial GEC, the discussion of the quantity $\omega(\beta, p^0)$, and the proof sketches of main theorems.

**Examples.** We call the class with a low $d_{\mathrm{GEC}}$ the *low self-play GEC class* and *low adversarial GEC class*. Next, we analyze the relation between the proposed classes and the following MG classes. We also propose a new decodable POMG class generalized from the single-agent POMDP. We note that except for the linear MG, the other classes cannot be analyzed by the recent posterior sampling work [XZS+22b]. We defer detailed definitions and proofs to Appendix B.

- **Linear MG.** This FOMG class admits a linear structure of the reward and transition by feature vectors $\boldsymbol{\phi}(s,a,b) \in \mathbb{R}^d$ as $r_h(s,a,b) = \mathbf{w}_h^\top \boldsymbol{\phi}(s,a,b)$ and $\mathbb{P}_h(s'|s,a,b) = \boldsymbol{\theta}_h(s')^\top \boldsymbol{\phi}(s,a,b)$ [XCWY20]. We then prove that "*linear MG $\subset$ low self-play/adversarial GEC*" with $d_{\mathrm{GEC}} = \widetilde{O}(H^3 d)$.

- **Linear Mixture MG.** This FOMG class admits a different type of the linear structure for the transition [CZG22] as $\mathbb{P}_h(s'|s,a,b) = \boldsymbol{\theta}_h^\top \boldsymbol{\phi}(s,a,b,s')$ with $\boldsymbol{\phi}(s,a,b,s') \in \mathbb{R}^d$. We prove that "*linear mixture MG $\subset$ low self-play/adversarial GEC*" with $d_{\mathrm{GEC}} = \widetilde{O}(H^3 d)$.

- **Low Self-Play Witness Rank.** [HLWY21] defines this FOMG class for self-play by supposing that an inner product of specific vectors in $\mathbb{R}^d$ defined on current models can lower bound witnessed model misfit and upper bound the Bellman error with a coefficient $\kappa_{\mathrm{wit}}$, which generalizes linear/linear mixture MGs. We can prove "*low self-play witness rank $\subset$ low self-play GEC*" with $d_{\mathrm{GEC}} = \widetilde{O}(H^3 d/\kappa_{\mathrm{wit}}^2)$.

- **$\alpha$-Weakly Revealing POMG.** This POMG class assumes $\min_h \sigma_S(\mathbb{O}_h) \geq \alpha$ where $\mathbb{O}_h \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{S}|}$ is the matrix by $\mathbb{O}_h(\cdot|\cdot)$ and $\sigma_S$ is the $S$-th singular value [LSJ22]. We prove that "*$\alpha$-weakly revealing POMG $\subset$ low self-play/adversarial GEC*" with $d_{\mathrm{GEC}} = \widetilde{O}(H^3|\mathcal{O}|^3|\mathcal{A}|^2|\mathcal{B}|^2|\mathcal{S}|^2/\alpha^2)$.

- **Decodable POMG.** We propose decodable POMGs by generalizing decodable POMDPs [EJKM22, DKJ$^+$19], assuming that an unknown decoder $\phi_h$ recovers states from observations, i.e., $\phi_h(o) = s$. We can prove "*decodable POMG $\subset$ low self-play/adversarial GEC*" with $d_{\mathrm{GEC}} = \widetilde{O}(H^3|\mathcal{O}|^3|\mathcal{A}|^2|\mathcal{B}|^2)$.

**Discussion of $\omega(\beta, p^0)$.** We briefly discuss the upper bound of the quantity $\omega(\beta, p^0)$ for FOMGs and POMGs. We refer readers to Appendix C for more detailed proofs. For FOMGs, according to Lemma 2 of [AZ22a], when $\mathcal{F}$ is finite, $p^0 = \mathrm{Unif}(\mathcal{F})$, then $\omega(\beta, p^0) \leq \log|\mathcal{F}|$ by its definition. When $\mathcal{F}$ is infinite, it shows that under mild conditions, there exists a prior $p^0$ over $\mathcal{F}$, $B \geq \log(6B^2/\epsilon)$, and $\nu = \epsilon/(6\log(6B^2/\epsilon))$ such that $\omega(\beta, p^0) \leq \beta\epsilon + \log(\mathcal{N}(\frac{\epsilon}{6\log(B/\nu)}))$, where $\mathcal{N}(\epsilon)$ is the $\epsilon$-covering number w.r.t. the distance $d(f,f') := \sup_{s,a,b,h} |D_{\mathrm{He}}^2(\mathbb{P}_{f,h}(\cdot\,|\,s,a,b), \mathbb{P}_{f^*,h}(\cdot\,|\,s,a,b)) - D_{\mathrm{He}}^2(\mathbb{P}_{f',h}(\cdot\,|\,s,a,b), \mathbb{P}_{f^*,h}(\cdot\,|\,s,a,b))|$. Since we have $|D_{\mathrm{He}}^2(P,R) - D_{\mathrm{He}}^2(Q,R)| \leq \frac{\sqrt{2}}{2}\|P-Q\|_1$ for any distributions $P, Q$, and $R$, the covering number w.r.t. the distance $d$ can connect to the more common covering number w.r.t. the $\ell_1$ distance. Thus, the upper bound of $\omega(\beta, p^0)$ can be calculated for different cases. Additionally, for POMGs, inspired by [AZ22a], our work proves that under similar conditions, $\omega(\beta, p^0)$ with finite and infinite $\mathcal{F}$ admit the same bounds as those for FOMGs. The difference is that the covering number is w.r.t. the distance $d(f,f') = \sup_{\pi,\nu} |D_{\mathrm{He}}^2(\mathbf{P}_{f,H}^{\pi,\nu}, \mathbf{P}_{f^*,H}^{\pi,\nu}) - D_{\mathrm{He}}^2(\mathbf{P}_{f',H}^{\pi,\nu}, \mathbf{P}_{f^*,H}^{\pi,\nu})|$, which further connects to the $\ell_1$ distance defined as $d_1(f,f') := \sup_{\pi,\nu} \|\mathbf{P}_{f,H}^{\pi,\nu} - \mathbf{P}_{f',H}^{\pi,\nu}\|_1$. Such a covering number under $\ell_1$ distance is further analyzed in [ZUSL22]. Our work gives the first detailed proof for the upper bound of $\omega(\beta, p^0)$ under the partially observable setting, which is thus of independent interest.

Next, we outline our proof sketches. Detailed proofs are deferred to Appendices D, E, and F.

**Proof Sketch of Theorem 3.1.** To prove Theorem 3.1, we only need to combine the result in Propositions 1 and 2 via $\mathbb{E}[\mathrm{Reg}^{\mathrm{sp}}(T)] = \mathbb{E}[\mathrm{Reg}_1^{\mathrm{sp}}(T) + \mathrm{Reg}_2^{\mathrm{sp}}(T)]$. We thus first give a proof sketch for Proposition 1. We decompose $\mathrm{Reg}_1^{\mathrm{sp}}(T) = \mathrm{Term(i)} + \mathrm{Term(ii)}$ where

$$\mathrm{Term(i)} = \textstyle\sum_{t=1}^T \left[V_{f^*}^* - V_{f^*}^{\pi^t,\underline{\nu}^t}\right], \quad \mathrm{Term(ii)} = \textstyle\sum_{t=1}^T \left[V_{f^*}^{\pi^t,\underline{\nu}^t} - V_{f^*}^{\pi^t,*}\right]$$

Intuitively, $\mathbb{E}[\mathrm{Term(i)}]$ is the main player's regret incurred Line 4 of Algorithm 1 and $\mathbb{E}[\mathrm{Term(ii)}]$ is the exploiter's regret incurred by Line 6. We further show

$$\mathrm{Term(i)} \leq \textstyle\sum_{t=1}^T \big[ - \Delta V_{\overline{f}^t}^* + V_{\overline{f}^t}^{\pi^t, \underline{\nu}^t} - V_{f^*}^{\pi^t, \underline{\nu}^t}\big], \quad \mathrm{Term(ii)} = \textstyle\sum_{t=1}^T \big[V_{f^*}^{\pi^t, \underline{\nu}^t} - V_{\underline{f}^t}^{\pi^t, \underline{\nu}^t} + \Delta V_{\underline{f}^t}^{\pi^t, *}\big]$$

where $\Delta V_{\overline{f}^t}^* := V_{\overline{f}^t}^{\pi^t, \overline{\nu}^t} - V_{f^*}^*$ and $\Delta V_{\underline{f}^t}^{\pi^t,*} = V_{\underline{f}^t}^{\pi^t, \underline{\nu}^t} - V_{f^*}^{\pi^t,*}$ are associated with the optimism terms in posterior distributions. The inequality above for Term(i) is due to Line 4 such that $V_{\overline{f}^t}^{\pi^t, \overline{\nu}^t} = \min_\nu V_{\overline{f}^t}^{\pi^t, \nu} \leq V_{\overline{f}^t}^{\pi^t, \underline{\nu}^t}$. By Definition 1 for self-play GEC, we obtain that $\sum_{t=1}^T \big(V_{\overline{f}^t, 1}^{\pi^t, \underline{\nu}^t} - V_{f^*}^{\pi^t, \underline{\nu}^t}\big)$ and $\sum_{t=1}^T \big(V_{f^*}^{\pi^t, \underline{\nu}^t} - V_{\underline{f}^t, 1}^{\pi^t, \underline{\nu}^t}\big)$ can be bounded by

$$\Big[d_{\mathrm{GEC}} \textstyle\sum_{h=1}^H \sum_{t=1}^T \Big( \sum_{\iota=1}^{t-1} \mathbb{E}_{(\sigma_{\exp}^\iota, h)} \ell(\rho^t, \xi_h^\iota)\Big)\Big]^{1/2} + 2H(d_{\mathrm{GEC}} HT)^{\frac{1}{2}} + \epsilon HT$$

where $\rho^t$ is chosen as $\overline{f}^t$ or $\underline{f}^t$ respectively. By Lemma 9 and Lemma 10, we prove that for both FOMGs and POMGs, the accumulation of the losses $\ell(\overline{f}^t, \xi_h^\iota)$ in (3) connects to the likelihood function $L_h^t$ defined in (1) and (2). Thus, we obtain $\mathbb{E}[\mathrm{Term(i)}] \leq \sum_{t=1}^T \mathbb{E}_{Z^{t-1}} \mathbb{E}_{\overline{f}^t \sim p^t} \{-\gamma_1 \Delta V_{\overline{f}^t}^* - \sum_{h=1}^H \sum_{\iota=1}^{t-1} [L_h^t(\overline{f}^t) - L_h^t(f^*)] + \log \frac{p^t(\overline{f}^t)}{p^0(\overline{f}^t)}\} + 2H(d_{\mathrm{GEC}} HT)^{\frac{1}{2}} + \epsilon HT$ and $\mathbb{E}[\mathrm{Term(ii)}]$ has a similar bound based on $q^t$, where $Z^{t-1}$ is the randomness history. By Lemma 8, the posterior distributions $p^t$ and $q^t$ following Lines 3 and 5 of Algorithm 1 can minimize the above upper bounds for $\mathbb{E}[\mathrm{Term(i)}]$ and $\mathbb{E}[\mathrm{Term(ii)}]$. Therefore, we can relax $p^t$ and $q^t$ to be distributions defined around the true model $f^*$ to enlarge above bounds. When $T$ is sufficiently large and $\eta = 1/2$, we have

$$\mathbb{E}[\mathrm{Term(i)}] \leq \omega(HT, p^0)T/\gamma_1 + \gamma_1 d_{\mathrm{GEC}} H/4 + 2H(d_{\mathrm{GEC}} HT)^{\frac{1}{2}} + \epsilon HT,$$
$$\mathbb{E}[\mathrm{Term(ii)}] \leq \omega(HT, q^0)T/\gamma_2 + \gamma_2 d_{\mathrm{GEC}} H/4 + 2H(d_{\mathrm{GEC}} HT)^{\frac{1}{2}} + \epsilon HT$$

Choosing proper values for $\epsilon$,$\gamma_1$, and $\gamma_2$, we obtain the bound for $\mathbb{E}[\mathrm{Reg}_1^{\mathrm{sp}}(T)]$ in Theorem 3.1 via $\mathrm{Reg}_1^{\mathrm{sp}}(T) = \mathrm{Term(i)} + \mathrm{Term(ii)}$. In addition, we can prove the bound of $\mathbb{E}[\mathrm{Reg}_2^{\mathrm{sp}}(T)]$ in a symmetric manner. Finally, combining $\mathbb{E}[\mathrm{Reg}_1^{\mathrm{sp}}(T)]$ and $\mathbb{E}[\mathrm{Reg}_2^{\mathrm{sp}}(T)]$ gives the result in Theorem 3.1.

**Proof Sketch of Theorem 4.1.** Under the adversarial setting, the policy of the opponent $\nu^t$ is not generated by the algorithm, which could be arbitrarily time-varying. We decompose $\mathrm{Reg}^{\mathrm{adv}}(T) = \sum_{t=1}^T \Delta V_{f^t}^* + \sum_{t=1}^T [V_{f^t}^* - V_{f^*}^{\pi^t, \nu^t}]$ where $\Delta V_{f^t}^* := V_{f^*}^* - V_{f^t}^*$ relates to optimism. Since $(\pi^t, \overline{\nu}^t)$ is NE of $V_{f^t}^{\pi, \nu}$ as in Line 3 of Algorithm 2, we have $V_{f^t}^* = \min_\nu V_{f^t}^{\pi^t, \nu} \leq V_{f^t}^{\pi^t, \nu^t}$, which leads to

$$\mathrm{Reg}^{\mathrm{adv}}(T) \leq \textstyle\sum_{t=1}^T \Delta V_{f^t}^* + \sum_{t=1}^T \big[V_{f^t}^{\pi^t, \nu^t} - V_{f^*}^{\pi^t, \nu^t}\big]$$

We can bound $\sum_{t=1}^T [V_{f^t}^{\pi^t, \nu^t} - V_{f^*}^{\pi^t, \nu^t}]$ via adversarial GEC in Definition 3 by

$$\Big[d_{\mathrm{GEC}} \textstyle\sum_{h=1}^H \sum_{t=1}^T \Big( \sum_{\iota=1}^{t-1} \mathbb{E}_{(\sigma_{\exp}^\iota, h)} \ell(f^t, \xi_h^\iota)\Big)\Big]^{\frac{1}{2}} + 2H(d_{\mathrm{GEC}} HT)^{\frac{1}{2}} + \epsilon HT$$

Connecting the loss $\ell(\overline{f}^t, \xi_h^\iota)$ to the likelihood function $L_h^t$ defined in (1) and (2) via Lemmas 9 and 10, we obtain $\mathbb{E}[\mathrm{Reg}^{\mathrm{adv}}(T)] \leq \sum_{t=1}^T \mathbb{E}_{Z^{t-1}} \mathbb{E}_{f^t \sim p^t} \{\gamma \sum_{t=1}^T \Delta V_{f^t}^* - \sum_{h=1}^H \sum_{\iota=1}^{t-1} [L_h^t(f^t) - L_h^t(f^*)] +$

$\log \frac{p^t(f^t)}{p^0(f^t)}\} + 2H(d_{\text{GEC}}HT)^{\frac{1}{2}} + \epsilon HT$. Lemma 8 shows $p^t$ in Line 3 of Algorithm 2 can minimize this bound. Thus, relaxing $p^t$ to be distribution defined around the true model $f^*$, with sufficiently large $T$ and $\eta = 1/2$, we have

$$\mathbb{E}[\text{Reg}^{\text{adv}}(T)] \leq \omega(4HT, p^0)T/\gamma + \gamma d_{\text{GEC}}H/4 + 2H(d_{\text{GEC}}HT)^{\frac{1}{2}} + \epsilon HT$$

Choosing proper values for $\epsilon$ and $\gamma$, we eventually obtain the bound for $\mathbb{E}[\text{Reg}^{\text{adv}}(T)]$ in Theorem 4.1.

# 6 Conclusion

This paper investigates posterior sampling algorithms for competitive reinforcement learning (RL) under general function approximation. We focus on two essential settings in zero-sum Markov games (MGs): self-play and adversarial learning, both of which are complicated by the challenge of partial observability.

We first introduced two novel complexity measures—the Self-Play and Adversarial Generalized Eluder Coefficients (GEC)—which capture the exploration-exploitation trade-off in these games. Using the Self-Play GEC, we developed a model-based posterior sampling method to efficiently control both players and learn Nash equilibria, even when the system is partially observable. For adversarial MGs, we incorporated the Adversarial GEC to design a model-based posterior sampling algorithm that effectively handles adversarial policies, while ensuring low regret bounds that scale sublinearly with both the GEC and the number of episodes.

Our approach is the first to provide a general framework for sample-efficient, model-based posterior sampling in competitive RL, applicable to a broad class of MGs under both full and partial observability. These results pave the way for more efficient exploration strategies in competitive multi-agent systems. Future research could explore extending this framework to broader multi-agent RL settings and more complex function approximation models.

# References

[AJ17] Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

[AJS+20] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[AZ22a] Alekh Agarwal and Tong Zhang. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*, 2022.

[AZ22b] Alekh Agarwal and Tong Zhang. Non-linear reinforcement learning in large action spaces: Structural conditions and sample-efficiency of posterior sampling. *arXiv preprint arXiv:2203.08248*, 2022.

[BBC+19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[BJ20] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

[BJY20] Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

[BLZZ23] Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*, 2023.

[BS19] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[CBM22] Fan Chen, Yu Bai, and Song Mei. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms, 2022.

[CCML18] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

[CD22] Qiwen Cui and Simon S Du. When is offline two-player zero-sum markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022.

[CL11] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

[CLY$^+$22] Zixiang Chen, Chris Junchi Li, Angela Yuan, Quanquan Gu, and Michael I Jordan. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

[CMB22] Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022.

[CYJW20] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[CZG22] Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *International Conference on Algorithmic Learning Theory*, pages 227–261. PMLR, 2022.

[DGZ22] Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

[DKJ$^+$19] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

[DKL$^+$21] Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

[DMZZ21] Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051, 2021.

[EJKM22] Yonathan Efroni, Chi Jin, Akshay Krishnamurthy, and Sobhan Miryoosefi. Provable reinforcement learning with a short-term memory. In *International Conference on Machine Learning*, pages 5832–5850. PMLR, 2022.

[FKQR21] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[GDB16] Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*, pages 510–518. PMLR, 2016.

[GMR22a] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable pomdps, without computationally intractable oracles. *arXiv preprint arXiv:2206.03446*, 2022.

[GMR22b] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022.

[HLWY21] Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

[JKA$^+$17] 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.

[JKKL20] Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

[JLM21] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

[JLWY21] Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

[JLY21] Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021.

[JMS22] Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games. *arXiv preprint arXiv:2204.04186*, 2022.

[JYWJ20] 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. PMLR, 2020.

[KAL16] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

[KECM21] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34:24523–24534, 2021.

[LCSJ22] Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pages 5175–5220. PMLR, 2022.

[LNSJ22] Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*, 2022.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

[LSJ22] Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable markov games. *arXiv preprint arXiv:2206.01315*, 2022.

[LVR17] Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

[LWJ22] Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. *arXiv preprint arXiv:2203.06803*, 2022.

[LYBJ21] Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

[MB21] Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum Markov games. *arXiv preprint arXiv:2110.05682*, 2021.

[NKLK20] Anusha Nagabandi, Kurt Konolige, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. In *Conference on Robot Learning*, pages 1101–1112. PMLR, 2020.

[NSZ$^+$22] Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. *arXiv preprint arXiv:2210.16976*, 2022.

[OBPVR16] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

[OVR14] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

[QWB$^+$22] Shuang Qiu, Lingxiao Wang, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, pages 18168–18210. PMLR, 2022.

[QWY$^+$21] Shuang Qiu, Xiaohan Wei, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. Provably efficient fictitious play policy optimization for zero-sum markov games with structured transitions. In *International Conference on Machine Learning*, pages 8715–8725. PMLR, 2021.

[QYWY21] Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free rl with kernel and neural function approximations: Single-agent mdp and markov game. In *International Conference on Machine Learning*, pages 8737–8747. PMLR, 2021.

[Rus19] Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

[RVR14] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[Sha53] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[SHM$^+$16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[SJK$^+$19] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

[SMB21] Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum Markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

[SSSS16] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[SV16] Igal Sason and Sergio Verdú. $f$-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[TWYS21] Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In *International conference on machine learning*, pages 10279–10288. PMLR, 2021.

[USL$^+$22a] Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Computationally efficient pac rl in pomdps with latent determinism and conditional embeddings. *arXiv preprint arXiv:2206.12081*, 2022.

[USL$^+$22b] Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*, 2022.

[VEB$^+$17] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. StarCraft II: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

[VH14] Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

[WCYW] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Represent to control partially observed systems: Representation learning with provable sample efficiency. In *The Eleventh International Conference on Learning Representations*.

[WSY20a] Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

[WSY20b] 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*, 33:6123–6135, 2020.

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[WWDK19] 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.

[XCWY20] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

[XZS$^+$22a] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

[XZS$^+$22b] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang. A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning*, pages 24496–24523. PMLR, 2022.

[YJW$^+$20] Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

[YZW$^+$23] Yunchang Yang, Han Zhong, Tianhao Wu, Bin Liu, Liwei Wang, and Simon S Du. A reduction-based framework for sequential decision making with delayed feedback. *arXiv preprint arXiv:2302.01477*, 2023.

[ZBB$^+$20] Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[ZGS21] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[Zha06] Tong Zhang. From $\varepsilon$-entropy to kl-entropy: Analysis of minimum information complexity density estimation. 2006.

[Zha22] Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

[ZLY22] Wenhao Zhan, Jason D Lee, and Zhuoran Yang. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. *arXiv preprint arXiv:2206.01588*, 2022.

[ZRY$^+$22] Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pages 26447–26466. PMLR, 2022.

[ZUSL22] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

[ZXT$^+$22] Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. *arXiv preprint arXiv:2202.07511*, 2022.

[ZXZ$^+$22] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. Gec: A unified framework for interactive decision making in mdp, pomdp, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

[ZYWJ23] Han Zhong, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Can reinforcement learning find stackelberg-nash equilibria in general-sum markov games with myopically rational followers? *Journal of Machine Learning Research*, 24(35):1–52, 2023.