Artificial Intelligence 348 (2025) 104392

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

Provably efficient information-directed sampling algorithms for multi-agent reinforcement learning

Qiaosheng Zhang [a,d], Chenjia Bai [b], Shuyue Hu [a], Zhen Wang [c,*], Xuelong Li [b,*]

[a] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[b] Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd, China
[c] Northwestern Polytechnical University, Xi'an, China
[d] Shanghai Innovation Institute, Shanghai, China

ARTICLE INFO

Keywords:
Multi-agent reinforcement learning
Markov games
Sample-efficient algorithms
Posterior sampling
Information theory
Rate-distortion theory

ABSTRACT

This work designs and analyzes a novel set of algorithms for multi-agent reinforcement learning (MARL) based on the principle of information-directed sampling (IDS). These algorithms draw inspiration from foundational concepts in information theory, and are proven to be sample efficient in MARL settings such as two-player zero-sum Markov games (MGs) and multi-player general-sum MGs. For episodic two-player zero-sum MGs, we present three sample-efficient algorithms for learning Nash equilibrium. The basic algorithm, referred to as MAIDS, employs an asymmetric learning structure where the max-player first solves a minimax optimization problem based on the *joint information ratio* of the joint policy, and the min-player then minimizes the *marginal information ratio* with the max-player's policy fixed. Theoretical analyses show that it achieves a Bayesian regret of $\tilde{O}(\sqrt{K})$ for K episodes. To reduce the computational load of MAIDS, we develop an improved algorithm called REG-MAIDS, which has the same Bayesian regret bound while enjoying less computational complexity. Moreover, by leveraging the flexibility of IDS principle in choosing the learning target, we propose two methods for constructing compressed environments based on rate-distortion theory, upon which we develop an algorithm COMPRESSED-MAIDS wherein the learning target is a compressed environment. Finally, we extend REG-MAIDS to multi-player general-sum MGs and prove that it can learn either the Nash equilibrium or coarse correlated equilibrium in a sample-efficient manner.

1. Introduction

The problem of multi-agent reinforcement learning (MARL), where multiple agents learn and make sequential decisions in a shared environment to optimize their individual or collective rewards, has become increasingly relevant in real-world applications such as robot systems [1], autonomous driving [2], multi-player games [3], etc. A crucial consideration in MARL is the *sample efficiency*, as it directly influences the practical applicability and scalability of MARL algorithms in real-world multi-agent systems. The pursuit of sample efficient algorithms stands as a critical and pressing concern in the realm of MARL.

* Corresponding authors.
E-mail addresses: zhangqiaosheng@pjlab.org.cn (Q. Zhang), baichenjia@pjlab.org.cn (C. Bai), hushuyue@pjlab.org.cn (S. Hu), w-zhen@nwpu.edu.cn (Z. Wang), xuelong_li@ieee.org (X. Li).

https://doi.org/10.1016/j.artint.2025.104392
Received 25 May 2024; Received in revised form 25 June 2025; Accepted 27 June 2025

Available online 10 July 2025

Table 1
A summary of theoretical guarantees for various MARL algorithms; some are provided as regret bounds (either frequentist regret or Bayesian regret), while others are given as probably approximately correct (PAC) sample complexity.

Settings	Algorithms	Regret	PAC
Two-player zero-sum MG	VI-explore [24]	$\tilde{O}((H^5S^2ABK^2)^{1/3})$, Frequentist	-
	VI-ULCB [24]	$\tilde{O}(\sqrt{H^4S^2ABK})$, Frequentist	$\tilde{O}(H^4S^2AB/\epsilon^2)$
	Nash Q-learning [4]	-	$\tilde{O}(H^5SAB/\epsilon^2)$
	Nash V-learning [4]	-	$\tilde{O}(H^6S(A+B)/\epsilon^2)$
	Nash VI [7]	-	$\tilde{O}(H^3SAB/\epsilon^2)$
	OMVI-SM [25]	$\tilde{O}(\sqrt{H^4S^3A^3B^3K})$, Frequentist	$\tilde{O}(H^4S^3A^3B^3/\epsilon^2)$
	Posterior sampling [8]	$\tilde{O}(\sqrt{H^4S^2A^2B^2K})$, Frequentist	-
	MAIDS/Reg-MAIDS (Theorems 1-3)	$\tilde{O}(\sqrt{H^4S^2ABK})$, Bayesian (for large K) $\tilde{O}(\sqrt{H^4S^3A^2B^2K})$, Bayesian (for any K)	-
Multi-player general-sum MG	Multi-Nash-VI [7]	$\tilde{O}(\sqrt{H^4S^2(\prod_{i=1}^N A_i)K})$, Frequentist (Nash, CE, CCE)	$\tilde{O}(H^4S^2(\prod_{i=1}^N A_i)/\epsilon^2)$ (Nash, CE, CCE)
	V-learning OMD [26]	$\tilde{O}(\sqrt{H^6S(\max_{i\in[N]} A_i)K})$, Frequentist (CCE)	$\tilde{O}(H^6S\max_{i\in[N]} A_i/\epsilon^2)$ (CCE)
	CCE-V-Learning [27]	-	$\tilde{O}(H^5S\max_{i\in[N]} A_i/\epsilon^2)$ (CCE)
	General-MAIDS (Theorem 5)	$\tilde{O}(\sqrt{H^4S^3(\prod_{i=1}^N A_i)^2K})$, Bayesian (Nash, CCE)	-

In recent years, there have been efforts to develop sample-efficient MARL algorithms, and some accompanied with theoretical guarantees. Most existing provably efficient algorithms are based on the principle of *optimism in the face of uncertainty* (OFU). Notable OFU-based algorithms include Nash Q-learning [4], Nash V-learning [5], and model-based algorithms [6,7]. Another principled but less explored algorithm design paradigm for MARL is *posterior sampling*. Very recently, two studies demonstrated that posterior sampling-based algorithms can be both sample efficient and computationally efficient in MARL, either with full observation [8] or partial observation [9].

In the general paradigm of posterior sampling, *information-directed sampling* (IDS) stands out as a relatively new yet principled exploration strategy for sequential decision-making problems [10–12]. Drawing inspiration from information theory, IDS tackles the exploration-exploitation tradeoff by requiring the agent to balance the policy's sub-optimality (exploitation) and the acquired information about a *learning target* (exploration) via a quantity called *information ratio*. Compared with OFU-based and other posterior sampling algorithms (such as Thompson sampling [13]), IDS has several appealing features. For instance, (i) it offers flexibility in choosing which kind of information to learn (referred to as the *learning target*) when encountering different information structures[1]; and (ii) IDS is able to accommodate parametric uncertainty and heteroscedastic observation noise (e.g., when states can only be partially observed) [14]. As a result, IDS outperforms Thompson sampling in terms of sample efficiency in a variety of settings [15]. Empirically, IDS-based algorithms have also demonstrated superior performance and computation efficiency in bandit problems [16, 17] and reinforcement learning (RL) problems [14,18], affirming its practicality in real-world scenarios.

Despite the theoretical advantages and empirical success, it remains unknown if the IDS principle can be applied to *competitive* or *cooperative* multi-player decision-making problems. If it can, the question that arises is whether IDS-based algorithms can maintain favorable features that have been demonstrated in bandit and RL problems. To address these unexplored questions, this work initiates an investigation into the potential of IDS in MARL settings, specifically focusing on the episodic two-player zero-sum Markov game (MG) and multi-player general-sum MG. We put forth the first line of MARL algorithms designed based on the IDS principle. One attractive feature of our algorithm is that when the environment (or transition kernel) admits a compressed approximation, such approximation can be utilized in constructing the learning target in our algorithm, resulting in an improved regret bound. The concept of compression is inspired by the classical lossy compression problems (a.k.a. rate-distortion theory) in information theory [19] and other related works in the RL literature [20–23].

1.1. Main contributions

This work introduces new IDS-based MARL algorithms that are driven by information theory. These algorithms enrich the set of sample-efficient MARL algorithms that are theoretically known to satisfy a $\tilde{O}(\sqrt{K})$ regret bound[2] for MGs with K episodes. A summary of theoretical guarantees for different MARL algorithms in this work and prior works is provided in Table 1. We detail our contributions as follows.

1. We first develop a basic self-play algorithm, named MAIDS, for learning Nash equilibrium (NE) in two-player zero-sum MGs. It operates by letting players sequentially optimize the *joint information ratio* and *marginal information ratio* at each episode. These

[1] This feature is particularly beneficial when the environment is overwhelmingly complex or when it consists of specific information structures that can be compressed. Example 1 illustrates how to leverage this feature to improve sample efficiency.

[2] We say $f(n) = \tilde{O}(g(n))$ if $f(n) = O(g(n) \cdot \text{polylog}(n))$.

ratios represent the expected regret over the acquired information about a learning target, and we choose the learning target to be the entire environment in MAIDS. Theorems 1 and 2 present Bayesian regret bounds that both scale as $\tilde{O}(\sqrt{K})$ for K episodes of MGs, and is valid for all prior distributions of the environment. The regret in Theorem 1 is comparatively loose, but it holds for all values of K. In contrast, Theorem 2 provides a tighter bound; however, it requires more assumptions and is only valid for sufficiently large K.
2. We also develop an algorithm REG-MAIDS that offers reduced computational complexity compared to MAIDS, without compromising the sample efficiency (Theorem 3). REG-MAIDS can be implemented efficiently by leveraging existing computationally efficient MARL algorithms.
3. Given the flexibility of the IDS principle in selecting the learning target, we have developed an algorithm named COMPRESSED-MAIDS, where the learning target is a *compressed environment* (instead of the entire environment in MAIDS). Inspired by lossy compression in information theory, we introduce two principles for constructing the compressed environment, and provide a Bayesian regret bound for COMPRESSED-MAIDS under a specific compressed environment (Theorem 4).
4. Finally, we extend REG-MAIDS to multi-player general-sum MGs, and show that the general algorithm GENERAL-MAIDS can learn either the NE or coarse correlated equilibrium (CCE) sample-efficiently through the derivation of Bayesian regret bounds (Theorem 5 and Theorem 6).

It is worth noting that while MGs can be viewed as extensions of Markov decision process (MDP) for RL, the unpredictability of other players' actions (called *non-stationary* in [28]) imposes new challenges in algorithm design and in the construction of compressed environments.

- In the design of IDS-based algorithms, we address the competitive nature of the MG by adopting an *asymmetric learning procedure*. Specifically, in order to learn an approximate NE policy for the max-player, the max-player first chooses a policy that optimizes the joint information ratio against a (fictitious) *worst-case opponent*, and the min-player subsequently optimizes the marginal information ratio to *assist* max-player's learning. This two-step procedure differs notably from previous IDS algorithms for bandit and single-agent RL settings. Additionally, our analytical techniques, which incorporate information-theoretic methods, also differ from the techniques for OFU-based and Thompson sampling algorithms in the context of MARL.
- In constructing compressed environments, we note that the method of partitioning the environment in the single-agent setting is not applicable to the multi-agent setting due to the unpredictability of opponents' actions. Instead, we introduce a new distortion measure between any pair of environment instances (see Eqn. (36)), which essentially quantifies the value difference between two instances under the worst-case policy that players may adopt. The introduction of this distortion measure also imposes new challenges in the analysis of our Compressed-MAIDS algorithm. Furthermore, we also put forth two methods for constructing compressed environments based on rate-distortion theory—soft compression and hard compression (see Section 6), while the prior work on single-agent setting [12] only considers a construction method similar to hard compression.

1.2. Related works

Driven by the successful applications of MARL techniques in real-world applications, there has been a growing focus on the theoretical exploration of MARL in recent years. A number of studies have put forth sample-efficient algorithms for various representative MARL settings, including the tabular zero-sum MG [24,4,29,7], zero-sum MG with linear function approximation [25,30] and with general function approximation [31,6,8,9,32,33], as well as general-sum MGs [34,35,27,36–42]. These works provide effective, efficient and theoretically-sound solutions for finding the (approximate) Nash Equilibrium of MGs, based on the principle of OFU or posterior sampling. In contrast, our proposed algorithms are founded on the IDS principle.

The IDS principle was first proposed by [10,11] for bandit settings, and has since been explored in both the bandit and broader RL settings [14,43,44,17,15]. Theoretical analyses of IDS in the RL setting are credited to [45] and [12]. The former develops a Bayesian regret bound when the prior distribution of the environment is specialized to the Dirichlet distribution, while the latter establishes prior-free Bayesian regret bounds for the first time. Although the regret bounds of the algorithms in [12] do not match the information-theoretic lower bound, we notice that, with the help of a new analytical technique for Thompson sampling [46], closing this gap becomes possible. Our work is closely related to [12], as the algorithm design and proof technique for MAIDS and REG-MAIDS are inspired by their work. However, substantial efforts are required to address the non-stationary nature of MGs. Furthermore, the construction of the compressed environment, as well as the design and analyses of COMPRESSED-MAIDS, are significantly different from theirs. Besides, [47] proposes a variant of IDS-based RL algorithm based on the use of *Stein information*, and demonstrates the advantage of computational efficiency of their algorithm through both theoretical and experimental analyses.

As the ultimate goal of MARL is to make good decisions rather than to estimate/learn the environment, it is often not necessary to learn/explore every granular detail of the environment. When the environment comprises redundant information that is not helpful for making decision, it can be counterproductive for agents to excessively focus on exploring/learning the redundant information rather than to exploit. This motivates the consideration of compressed environments, and coincidentally, the IDS principle allows for the selection of a compressed environment as the learning target. In recent RL literature, [12] proposes a novel approach for constructing a compressed environment based on the partition of transition kernels and value functions, whereas the subsequent work [46] provides a similar but refined construction. However, their approaches cannot be directly applied to MARL due to the effect of the opponent's unpredictable policy on the value functions. Meanwhile, [22] proposes a construction of the compressed environment based on the rate-distortion theory, where the distortion measure is defined through the *value equivalence principle* [48,49]. It is worth mentioning that our soft-compression principle is similar to their construction, but the accompanying distortion measure differs. A detailed comparison between the distortion measures is provided in Remark 2, Section 6.

Connections to DEC: A concept that is related to IDS is the *decision-estimation coefficient* (DEC), which is a complexity measure for sequential decision-making [50]. The DEC can be decomposed into two terms: one that represents the regret and another that quantifies the cumulative estimation error. The cumulative estimation error is measured by the squared Hellinger distance between the trajectories induced by the true model and induced by the estimated model. In contrast to the estimation error, our work focuses on how much information can be acquired through the interaction with the environment, where the amount of information is measured by a mutual information term. It is worth noting that there is a similarity between our mutual information term and their estimation error term, to some extent. Specifically, mutual information is equivalent to the KL-divergence between distributions of trajectories, while their estimation error is measured by the squared Hellinger distance between distributions of trajectories.

Besides, a recent work [38] generalizes the concept of DEC to multi-agent settings, and proposes a complexity measure called *multi-agent decision-estimation coefficient* (MA-DEC) for multi-agent decision making problems. Our work differs from [38] in three aspects: (i) They consider frequentist settings while we consider Bayesian settings; (ii) The MA-DEC depends on the estimation error about the environment, while the information ratio in our work depends on the acquired information about the environment (or about the compressed environment); (iii) They focus on the statistical complexity while we additionally provide a computationally efficient algorithm (i.e., the Reg-MAIDS).

Connections to AIR: The IDS principle is also related to the concept of *algorithmic information ratio* (AIR) [51]. Their work establishes a theory for analyzing frequentist regrets through Bayesian-type algorithms in sequential decision-making problems. A central object to be optimized in their work is the AIR, which, in addition to encompassing the terms of expected regret and acquired information involved in the information ratio term, further integrates a reference distribution to be aligned with. However, to the best of our knowledge, investigations into AIR are restricted to the simpler bandit and RL settings, while their applicability in the competitive MARL environment remains unknown.

1.3. Outline

The paper is organized as follows. Preliminaries on notational conventions and related information-theoretic concepts are introduced in Section 2. The mathematical formulation of the two-player zero-sum MG and the learning objectives are introduced in Section 3. We then present our sample-efficient algorithms MAIDS, Reg-MAIDS, and Compressed-MAIDS (designed for zero-sum MGs) in Sections 4-6, respectively. In Section 7, we introduce the mathematical formulation of the multi-player general-sum MG as well as our accompanying algorithms. Section 8 concludes this work and proposes future research directions. Most detailed proofs are provided in the Appendices.

2. Preliminaries

2.1. Notations

For any positive integer $n \in \mathbb{N}^+$, we denote the set of positive integers ranging from 1 to n by $[n] \triangleq \{1, 2, \dots, n\}$. For any set $\mathcal{X}$, let $\Delta(\mathcal{X})$ be the probability simplex over $\mathcal{X}$ (i.e., the set of all possible probability distributions on $\mathcal{X}$). For any probability measures P and Q on a same measurable space $\mathcal{X}$, we define their *KL-divergence* (a.k.a. *relative entropy*) as

$$\mathbb{D}_{\mathrm{KL}}(P\|Q) \triangleq \int_{\mathcal{X}} \log(P(dx)/Q(dx))P(dx)$$

if P is absolutely continuous with respect to Q, where $P(dx)/Q(dx)$ is the Radon–Nikodym derivative of P with respect to Q. Moreover, we define the *total variation distance* between P and Q as $\mathbb{D}_{\mathrm{TV}}(P,Q) \triangleq \sup_{A\subseteq\mathcal{X}} |P(A) - Q(A)|$. The Pinsker's inequality states that

$$\mathbb{D}_{\mathrm{TV}}(P,Q) \le \sqrt{\frac{1}{2}\mathbb{D}_{\mathrm{KL}}(P\|Q)}.$$

We adopt *asymptotic notations*, including $O(.)$, $o(.)$, $\Omega(.)$, $\omega(.)$, and $\Theta(.)$, to describe the limiting behavior of functions/sequences. For instance, we say a pair of functions $f(n)$ and $g(n)$ satisfies $f(n) = O(g(n))$ if there exist $m > 0$ and $N_0 \in \mathbb{N}^+$ such that for all $n > N_0$, $|f(n)| \le mg(n)$. Moreover, we say $f(n) = \tilde{O}(g(n))$ if $f(n) = O(g(n) \cdot \mathrm{polylog}(n))$.

2.2. Preliminaries on information theory

For a pair of random variables X and Y, we define their *mutual information* as

$$\mathbb{I}(X;Y) \triangleq \mathbb{D}_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot\,)\|\mathbb{P}(X \in \cdot\,) \times \mathbb{P}(Y \in \cdot\,)),$$

which is also equivalent to the form $\mathbb{E}_X[\mathbb{D}_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot\,|X)\|\mathbb{P}(Y \in \cdot\,)]$. When introducing another random variable Z, one can define the *conditional mutual information* as

$$\mathbb{I}(X;Y|Z) \triangleq \mathbb{E}_Z[\mathbb{D}_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot\,|Z)\|\mathbb{P}(X \in \cdot\,|Z) \times \mathbb{P}(Y \in \cdot\,|Z))].$$

For a collection of random variables $(X_0, X_1, X_2, \ldots, X_n)$, the *chain rule* of mutual information [52] states that

$$\mathbb{I}(X_0; X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \mathbb{I}(X_0; X_i | X_1, \ldots, X_{i-1}).$$

Lossy compression (a.k.a., rate-distortion theory) is a classical topic in information theory that addresses the problem of lossy data compression. Consider a random variable X supported on $\mathcal{X}$ with probability distribution P_X, and a distortion measure $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that quantifies the distance between two symbols belonging to $\mathcal{X}$. The *rate–distortion function* $R(D) \triangleq \min_{\hat{X}:\mathbb{E}[d(X,\hat{X})] \le D} \mathbb{I}(X; \hat{X})$ describes the minimum average number of bits needed to compress the random variable X with respect to the distortion constraint D, and the "compressed" random variable $\hat{X}$ that achieves $\arg\min_{\hat{X}:\mathbb{E}[d(X,\hat{X})] \le D} \mathbb{I}(X; \hat{X})$ is of particular interest. Moreover, in addition to considering the *average distortion constraint* in the form of $\mathbb{E}[d(X, \hat{X})] \le D$, one may also consider a more strict distortion constraint in the form of $\mathbb{P}(d(X, \hat{X}) > D) = 0$, which is also called the *zero-excess distortion constraint*, that requires the distortion does not exceed a specified threshold with probability one. Inspired by lossy compression, in this work, we construct compressed environments by first defining a suitable distortion measure in the context of Markov games, and then find random variables that satisfy either the average distortion constraint (named *soft-compression*) or zero-excess distortion constraint (named *hard-compression*). We refer the reader to Section 6 for more details.

3. Zero-sum Markov games

In this section, we first introduce the mathematical formulation of the zero-sum MG that comprises two competitive players. A summary of notations used in the zero-sum MG is provided in Appendix A.1. The formulation of the more general multi-player general-sum MG is provided in Section 7.

3.1. The basic model

The two-player zero-sum MG is denoted by $\mathcal{E} = (H, \mathcal{S}, \mathcal{A}, \mathcal{B}, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$, where H is the length of each episode, $\mathcal{S}$ is the set of countable state space with cardinality $|\mathcal{S}| = S$, while $\mathcal{A}$ and $\mathcal{B}$ are the sets of action spaces of the max-player and min-player respectively, with cardinalities $|\mathcal{A}| = A$ and $|\mathcal{B}| = B$. For each step $h \in [H]$, $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ is the *transition kernel* from the current state and actions to the next state. We use $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ to denote the deterministic reward function. Without loss of generality, it is assumed that $\mathcal{S}$, $\mathcal{A}$, $\mathcal{B}$, $\{r_h\}_{h=1}^H$ are known[3] while the transition kernels $\{P_h\}_{h=1}^H$ are unknown and random. We also refer to $\mathcal{E}$ as the *environment* of the MARL problem.

3.2. Prior distributions

We consider a Bayesian setting where we have a prior belief on the environment $\mathcal{E}$, or equivalently, on the transition kernel $\{P_h\}_{h=1}^H$, since the other model parameters $\mathcal{S}, \mathcal{A}, \mathcal{B}$ and $\{r_h\}$ are assumed to be deterministic. For each step $h \in [H]$, let Θ_h be the parameter space of P_h, and let ρ_h be the *prior distribution* of P_h on Θ_h. Let $\Theta \triangleq \prod_{h=1}^H \Theta_h$ be the parameter space of the kernels $\{P_h\}_{h=1}^H$, and without loss of generality we assume Θ is convex. Let $\rho \triangleq \prod_{h=1}^H \rho_h$ be the *product* prior distribution of $\{P_h\}_{h=1}^H$ on Θ. The players know the prior distribution ρ but not the realization of $\{P_h\}_{h=1}^H$. Note that the environment $\mathcal{E}$ can be viewed as a random variable with distribution ρ since its only randomness comes from $\{P_h\}_{h=1}^H$.

3.3. Interaction processes

In the zero-sum MG, each episode $k \in [K]$ starts at an initial state s_1^k. At each step $h \in [H]$, both the max-player and min-player observe the current state s_h^k, and pick their own actions a_h^k and b_h^k *simultaneously*. The max-player receives a reward $r_h^k = r_h(s_h^k, a_h^k, b_h^k)$, while the min-player receives $-r_h^k$. The environment then transits to the next state s_{h+1}^k according to the transition kernel $P_h(\cdot | s_h^k, a_h^k, b_h^k)$. The episode ends when the final state s_{H+1}^k is reached. Without loss of generality and for simplicity, we assume that the initial state of each episode s_1^k is fixed to the state s_1 over episodes. Generalizing such an assumption to having a random initial state with a fixed but known distribution will not pose significant challenges.

3.4. Policies

A policy μ of the max-player is a collection of mappings $(\mu_1, \ldots, \mu_H)$ such that $\mu_h : \Omega_{h-1} \times \mathcal{S} \to \Delta(\mathcal{A})$. Note that in the MG with simultaneous moves, the max-player cannot observe the min-player's action when choosing her/his own action. The policy of the min-player is a collection of mappings $\nu = (\nu_1, \ldots, \nu_H)$ such that $\nu_h : \Omega_{h-1} \times \mathcal{S} \to \Delta(\mathcal{B})$. Moreover, we say a policy is a *Markov policy* if each mapping $\mu_h : \mathcal{S} \to \Delta(\mathcal{A})$ (or $\nu_h : \mathcal{S} \to \Delta(\mathcal{B})$) only takes the current state as the input; that is, the action only depends on the current state but not the past trajectory. We denote the set of all Markov policies of the max-player by Π_A, and that of the min-player by Π_B. Without loss of generality, we only consider Markov policies in the remaining parts of the paper.

[3] Extending to unknown and stochastic reward functions will not pose significant challenges, as learning transition kernels is more difficult than learning reward functions.

3.5. Value functions

The goal of the max-player, as the name suggests, is to find a policy that maximizes the cumulative reward, while the goal of the min-player is to minimize the cumulative reward. The cumulative reward can be represented by the *value function* and the *action-value function* introduced below. The value function $V^{\mathcal{E}}_{h,\mu,\nu} : \mathcal{S} \to \mathbb{R}$ at step h with respect to the environment $\mathcal{E}$ and the policy (μ, ν) is defined as

$$V^{\mathcal{E}}_{h,\mu,\nu}(s) \triangleq \mathbb{E}^{\mathcal{E}}_{\mu,\nu}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})\Big| s_h = s\right], \quad \forall s \in \mathcal{S}, \tag{1}$$

where $\mathbb{E}^{\mathcal{E}}_{\mu,\nu}$ denotes the expectation over the trajectory $\{s'_h, a'_h, b'_h\}_{h'=h}^{H}$ that is generated by the interaction between the policy (μ, ν) and environment $\mathcal{E}$. The value function $V^{\mathcal{E}}_{h,\mu,\nu}(s)$ represents the expected cumulative reward from state s at step h when executing policy (μ, ν). We also define the action-value function $Q^{\mathcal{E}}_{h,\mu,\nu} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ at step h with respect to the environment $\mathcal{E}$ and the policy (μ, ν) as

$$Q^{\mathcal{E}}_{h,\mu,\nu}(s, a, b) \triangleq \mathbb{E}^{\mathcal{E}}_{\mu,\nu}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'})\Big| s_h = s, a_h = a, b_h = b\right], \quad \forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}. \tag{2}$$

The action-value function $Q^{\mathcal{E}}_{h,\mu,\nu}(s, a, b)$ represents the expected cumulative reward from state s at step h when executing actions $(a_h, b_h) = (a, b)$ at step h and policy (μ, ν) afterwards. Moreover, we have the following Bellman equations:

$$Q^{\mathcal{E}}_{h,\mu,\nu}(s, a, b) = r_h(s, a, b) + \mathbb{E}_{s' \sim P_h(\cdot|s,a,b)}[V^{\mathcal{E}}_{h+1,\mu,\nu}(s')], \tag{3}$$

$$V^{\mathcal{E}}_{h,\mu,\nu}(s) = \mathbb{E}_{a \sim \mu_h(\cdot|s), b \sim \nu_h(\cdot|s)}[Q^{\mathcal{E}}_{h,\mu,\nu}(s, a, b)]. \tag{4}$$

3.6. Best responses

For any Markov policy μ of the max-player, there exists a *best response* $\nu^\dagger(\mu)$ from the min-player, which is a Markov policy that satisfies $V^{\mathcal{E}}_{h,\mu,\nu^\dagger(\mu)}(s) = \inf_\nu V^{\mathcal{E}}_{h,\mu,\nu}(s)$ for all (s, h). Similarly, if the min-player's policy ν is given, the max-player can also find a best response $\mu^\dagger(\nu)$ that satisfies $V^{\mathcal{E}}_{h,\mu^\dagger(\nu),\nu}(s) = \sup_\mu V^{\mathcal{E}}_{h,\mu,\nu}(s)$ for all (s, h). For notational convenience, we introduce the following abbreviations of value functions:

$$V^{\mathcal{E}}_{h,\mu,\dagger}(s) \triangleq V^{\mathcal{E}}_{h,\mu,\nu^\dagger(\mu)}(s), \quad \text{and} \quad V^{\mathcal{E}}_{h,\dagger,\nu}(s) \triangleq V^{\mathcal{E}}_{h,\mu^\dagger(\nu),\nu}(s). \tag{5}$$

3.7. Nash equilibrium

It is well known [53] that for a MG represented by $\mathcal{E}$, there exists a *Nash Equilibrium (NE) policy* (μ^*, ν^*), where both μ^* and ν^* are Markov policies, that satisfies

$$V^{\mathcal{E}}_{h,\mu^*,\dagger}(s) = \sup_\mu V^{\mathcal{E}}_{h,\mu,\dagger}(s) \quad \text{and} \quad V^{\mathcal{E}}_{h,\dagger,\nu^*}(s) = \inf_\nu V^{\mathcal{E}}_{h,\dagger,\nu}(s), \quad \forall (s, h) \in \mathcal{S} \times [H]. \tag{6}$$

This means that the policy μ^* (or policy ν^*) is optimal when the opponent can always choose the best response, i.e., it can be regarded as "the best response to the best response". The NE policy (μ^*, ν^*) also satisfies the minimax equation:

$$\sup_\mu \inf_\nu V^{\mathcal{E}}_{h,\mu,\nu}(s) = V^{\mathcal{E}}_{h,\mu^*,\nu^*}(s) = \inf_\nu \sup_\mu V^{\mathcal{E}}_{h,\mu,\nu}(s) \quad \forall (s, h) \in \mathcal{S} \times [H]. \tag{7}$$

Moreover, for each player, there is no incentive to move away from its NE policy if the other player does not move, in the sense that

$$V^{\mathcal{E}}_{h,\mu,\nu^*}(s) \le V^{\mathcal{E}}_{h,\mu^*,\nu^*}(s) \le V^{\mathcal{E}}_{h,\mu^*,\nu}(s) \tag{8}$$

for any policies μ and ν and for all $(s, h) \in \mathcal{S} \times [H]$. We sometimes write (μ^*, ν^*) as $(\mu^*(\mathcal{E}), \nu^*(\mathcal{E}))$ and $V^{\mathcal{E}}_{h,\mu^*,\nu^*}(s)$ as $V^{\mathcal{E}}_{h,\mu^*(\mathcal{E}),\nu^*(\mathcal{E})}(s)$ to explicitly highlight the fact that the Nash policy (μ^*, ν^*) depends on the environment $\mathcal{E}$. It is also known that the NE policy may not be unique; however, the corresponding values are the same. Thus, one can abbreviate the value $V^{\mathcal{E}}_{h,\mu^*,\nu^*}(s)$ of any NE policy (μ^*, ν^*) as $V^{\mathcal{E},*}_h(s)$ for simplicity. We refer to $V^{\mathcal{E},*}_h(s)$ as the *Nash value*.

3.8. Learning objectives

We first focus on the max-player. The goal of the max-player is to learn a policy μ that is almost as good as the NE policy μ^*, in the sense that the corresponding value $V^{\mathcal{E}}_{1,\mu,\dagger}(s_1)$ against the best response of the min-player is close to the value $V^{\mathcal{E}}_{1,\mu^*,\dagger}(s_1)$. Note that $V^{\mathcal{E}}_{1,\mu^*,\dagger}(s_1)$ is equal to $V^{\mathcal{E},*}_1(s_1)$ by the definition of NE policy. Thus, an appropriate way to define the regret of the max-player is through the difference between $V^{\mathcal{E},*}_1(s_1)$ and $V^{\mathcal{E}}_{1,\mu,\dagger}(s_1)$.

Suppose the two players interact with the environment for $K \in \mathbb{N}^+$ episodes. Let $\mu = \{\mu^k\}_{k=1}^K$ be the max-player's policy, where μ_k is the policy for episode k. For a fixed realization of the environment[4] $\mathcal{E} = e$, the cumulative regret over K episodes is defined as

$$\mathrm{Reg}_K(e,\mu) \triangleq \sum_{k=1}^{K} V_1^{e,*}(s_1) - V^e_{1,\mu_k,\dagger}(s_1). \tag{9}$$

The Bayesian regret is defined as

$$\mathrm{BR}_K(\mu) \triangleq \mathbb{E}\left(\mathrm{Reg}_K(\mathcal{E},\mu)\right) = \mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu_k,\dagger}(s_1)\right), \tag{10}$$

where the expectation is over the randomness in the environment $\mathcal{E}$ (or equivalently, over the prior probability measure ρ of the transition kernels $\{P_h\}_{h=1}^H$). Note that one can also define the regret $\mathrm{Reg}_K(e,\nu)$ and Bayesian regret $\mathrm{BR}_K(\nu)$ for the min-player in a symmetric fashion. Adding the (Bayesian) regrets of the two players together yields the classical duality gap $\sum_k (V^e_{1,\dagger,\nu^k} - V^e_{1,\mu^k,\dagger})$ and its Bayesian version. A zero duality gap implies NE.

4. The basic algorithm: MAIDS

This section considers zero-sum MGs and presents a basic multi-agent version of IDS-based algorithm, named MAIDS, with theoretical guarantees on Bayesian regrets.

Let $\mathcal{T}^k_{H+1} \triangleq \{s^k_1, a^k_1, b^k_1, r^k_1, \ldots, s^k_H, a^k_H, b^k_H, r^k_H, s^k_{H+1}\}$ be the *trajectory* of episode k, and $\mathcal{D}_k \triangleq \{\mathcal{T}^i_{H+1}\}_{i=1}^{k-1}$ be the full trajectory up to the *beginning* of episode k (with $\mathcal{D}_1 \triangleq \emptyset$). Recall that the environment $\mathcal{E}$ is random. In the following, let $\mathbb{E}_k(\cdot) \triangleq \mathbb{E}(\cdot|\mathcal{D}_k)$ denote the expectation w.r.t. the posterior distribution $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$, and let $\mathbb{I}_k(X;Y) \triangleq \mathbb{D}_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot\ |\mathcal{D}_k) \| \mathbb{P}(X \in \cdot\ |\mathcal{D}_k) \times \mathbb{P}(Y \in \cdot\ |\mathcal{D}_k))$ be the mutual information conditioned on $\mathcal{D}_k$. Note that $\mathcal{D}_k$ is also a random variable, thus $\mathbb{I}_k(X;Y)$ itself is a random variable. We also point out that $\mathbb{I}_k(X;Y)$ is different from the *conditional mutual information* $\mathbb{I}(X;Y|\mathcal{D}_k)$. In fact, we have $\mathbb{I}(X;Y|\mathcal{D}_k) = \mathbb{E}_{\mathcal{D}_k}(\mathbb{I}_k(X;Y))$, where $\mathbb{E}_{\mathcal{D}_k}(\cdot)$ denotes the expectation over the randomness of $\mathcal{D}_k$.

4.1. *Information ratio*

At the heart of the proposed algorithm is the notion of *joint information ratio*:

$$\Gamma_k(\mu,\nu,\chi) \triangleq \frac{(\mathbb{E}_k[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\mathcal{E}}_{1,\mu,\nu}(s_1)])^2}{\mathbb{I}^{\mu,\nu}_k(\chi;\mathcal{T}^k_{H+1})}, \tag{11}$$

which is defined for each episode $k \in [K]$. Here, χ is called the *learning target*, of which the most natural choice is the environment $\mathcal{E}$. The superscript μ,ν in $\mathbb{I}^{\mu,\nu}_k(\chi;\mathcal{T}^k_{H+1})$ means that the trajectory $\mathcal{T}^k_{H+1}$ is obtained by executing the policy (μ,ν). Note that the information ratio Γ_k explicitly depends on (μ,ν,χ), and also implicitly depends on the past trajectory $\mathcal{D}_k$ since both $\mathbb{E}_k[\cdot]$ and $\mathbb{I}^{\mu,\nu}_k(\cdot;\cdot)$ are calculated based on the random environment $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$. The numerator of Γ_k measures the (squared) expected difference between the values induced by policy $(\mu^*(\mathcal{E}),\nu)$ and by policy (μ,ν), which can be understood as the sub-optimality of μ for a fixed ν. The denominator measures the information about χ learned by the policy (μ,ν) through the interacted trajectory $\mathcal{T}^k_{H+1}$. Roughly speaking, the joint information ratio measures the *cost* of learning a unit of information about the learning target χ for the policy (μ,ν).

For a fixed max-player's policy μ, we further define the notion of *marginal information ratio* w.r.t. μ as

$$\Lambda^\mu_k(\nu,\chi) \triangleq \frac{(\mathbb{E}_k[V^{\mathcal{E}}_{1,\mu,\nu}(s_1) - V^{\mathcal{E}}_{1,\mu,\dagger}(s_1)])^2}{\mathbb{I}^{\mu,\nu}_k(\chi;\mathcal{T}^k_{H+1})}. \tag{12}$$

Here, the numerator measures the expected regret induced by the min-player's policy ν w.r.t. the fixed max-player's policy μ. Again, the marginal information ratio $\Lambda^\mu_k(\nu,\chi)$ implicitly depends on the past trajectory $\mathcal{D}_k$.

4.2. *Algorithm descriptions*

We now introduce our algorithm MAIDS, in which the learning target χ is selected as the environment $\mathcal{E}$. The pseudocode of MAIDS is provided in Algorithm 1. At the beginning of episode k, the max-player first calculates the posterior distribution of the environment $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$ based on the past trajectory $\mathcal{D}_k$ and the prior distribution ρ, and then chooses her policy as

$$\mu^k_{\mathrm{IDS}} = \underset{\mu\in\Pi_A}{\arg\min} \max_{\nu\in\Pi_B} \ \Gamma_k(\mu,\nu,\mathcal{E}), \tag{13}$$

[4] We prefer using the uppercase letter $\mathcal{E}$ to denote the random variable while using the lowercase letter e to denote a realization.

Algorithm 1 MAIDS for two-player zero-sum Markov Games.

1: **Input:** Prior distribution ρ
2: **for** episode $k = 1$ **to** K **do**
3: Both players calculate the posterior of environment $\mathcal{E}$ based on prior ρ and trajectory $\mathcal{D}_k$:

$$\mathbb{P}(\mathcal{E}|\mathcal{D}_k) \propto \rho(\mathcal{E}) \prod_{i=1}^{k-1} \prod_{h=1}^{H} P_h^{\mathcal{E}}(s_{h+1}^i | s_h^i, a_h^i, b_h^i). \tag{15}$$

4: The max-player calculates $\Gamma_k(\mu, \nu, \mathcal{E})$ for each μ, ν (where $\mathcal{E} \sim \mathbb{P}(\mathcal{E}|\mathcal{D}_k)$), and chooses

$$\mu_{\mathrm{IDS}}^k = \arg\min_{\mu \in \Pi_A} \max_{\nu \in \Pi_B} \Gamma_k(\mu, \nu, \mathcal{E}).$$

5: The min-player calculates $\Lambda_k^{\mu_{\mathrm{IDS}}^k}(\nu, \mathcal{E})$ for each ν (where $\mathcal{E} \sim \mathbb{P}(\mathcal{E}|\mathcal{D}_k)$), and chooses

$$\nu_{\mathrm{IDS}}^k = \arg\min_{\nu \in \Pi_B} \Lambda_k^{\mu_{\mathrm{IDS}}^k}(\nu, \mathcal{E}).$$

6: **end for**

implying that the max-player aims to minimize the joint information ratio by considering the presence of a worst-case opponent. This approach makes intuitive sense because the max-player has to make decisions prior to knowing the min-player's policy, so she should be more conservative.

The min-player then chooses her policy based on the knowledge of μ_{IDS}^k and the posterior distribution $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$:

$$\nu_{\mathrm{IDS}}^k = \arg\min_{\nu \in \Pi_B} \Lambda_k^{\mu_{\mathrm{IDS}}^k}(\nu, \mathcal{E}), \tag{14}$$

which minimizes the marginal information ratio w.r.t. the max-player's policy μ_{IDS}^k.

4.3. Regret bounds

First, we provide a non-asymptotic Bayesian regret bound for MAIDS that holds for any value of K.

Theorem 1. *Suppose the max-player's policy is $\mu_{IDS} = \{\mu_{IDS}^k\}_{k\in[K]}$ and the min-player's policy is $\nu_{IDS} = \{\nu_{IDS}^k\}_{k\in[K]}$, then for any prior distribution ρ, the Bayesian regret of μ_{IDS} satisfies*

$$\mathsf{BR}_K(\mu_{IDS}) \le 8\sqrt{H^4 S^3 A^2 B^2 K \log(SKH)}. \tag{16}$$

Proof of Theorem 1. See Section 4.4 for a proof sketch and Appendix B for the detailed proofs. ☐

Remark 1. In a symmetric fashion, the min-player can also run a MAIDS algorithm (tailored to the min-player) to obtain a policy $\tilde{\nu}_{\mathrm{IDS}}$ that has a bounded Bayesian regret $\mathsf{BR}_K(\tilde{\nu}_{\mathrm{IDS}})$ (as per Theorem 1), with the assistance of the max-player. The joint policy $(\mu_{\mathrm{IDS}}, \tilde{\nu}_{\mathrm{IDS}})$ of the two players, where μ_{IDS} is the one in Theorem 1, thus has a bounded Bayesian duality gap and is close to the NE policy. Such an asymmetric learning structure has also been adopted in [31,6,8]; however, adapting this trick for IDS requires different analysis techniques compared to prior works that are based on OFU or Thompson sampling.

Note that the regret bound is valid for all possible prior distribution ρ of the environment $\mathcal{E}$, and the scaling $\tilde{O}(\mathrm{poly}(S, A, B, H) \cdot \sqrt{K})$ is order-optimal w.r.t. the number of episodes K. Meanwhile, we note that the scalings of (S, A, B, H) have not yet matched the information-theoretic lower bound [24]. However, if we are given additional assumptions on *posterior consistency* and the boundedness of the policy sets Π_A and Π_B, and when the number of episodes K is sufficiently large, we can derive a more refined regret bound that reduces a factor of $\sqrt{SAB}$ compared to Theorem 1, by adopting the new proof techniques developed in [46] for Thompson sampling (TS). We denote the TS policy of the max-player as μ_{TS}^k. It first samples a realization of the environment $\mathcal{E} = e$ according to the distribution $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$, and then chooses the max-player's NE policy $\mu^*(e)$ with respect to e. Note that the TS policy μ_{TS}^k is a Markov policy.[5]

Assumption 1. (1) Let $|\Pi_A| < \infty$ and $|\Pi_B| < \infty$ be bounded. (2) There exists a strongly consistent estimator of the true environment given the collection of trajectories.

Theorem 2. *Suppose the max-player's policy is $\mu_{IDS} = \{\mu_{IDS}^k\}_{k\in[K]}$ and the min-player's policy is $\nu_{IDS} = \{\nu_{IDS}^k\}_{k\in[K]}$, and suppose Assumption 1 holds. For any prior distribution ρ, there exists a fixed positive constant K_0 such that for all $K > K_0$,*

$$\mathsf{BR}_K(\mu_{IDS}) \le 8\sqrt{H^4 S^2 ABK \log(SKH)} + K_0 H. \tag{17}$$

[5] Strictly speaking, the TS policy μ_{TS}^k is a Markov policy that incorporates an additional random seed for sampling an environment $\mathcal{E} = e$. For rigorousness, one needs to slightly modify the definition of the set of Markov policies Π_A to include all Markov policies with random seeds, thus encompassing the TS policy $\mu_{\mathrm{TS}}^k \in \Pi_A$. However, for ease of presentation, we maintain the definition of Π_A as presented in Section 3 of the main text, and such imprecision can be easily fixed.

While the regret bound in Theorem 2 improves by a factor of $\sqrt{SAB}$ compared to Theorem 1, it is only valid for sufficiently large K (since, according to Theorem 1, K must exceed K_0, which can be large). We defer the detailed proofs of Theorem 2 to Appendix C.

4.4. Proof sketch of Theorem 1

Note that for any policy $(\mu,\nu)=(\{\mu^k\}_{k\in[K]},\{\nu^k\}_{k\in[K]})$, the Bayesian regret over K episodes can be decomposed as

$$\begin{aligned}
&\mathsf{BR}_K(\mu)\\
&=\mathbb{E}\left(\sum_{k=1}^K V_1^{\mathcal{E},*}(s_1)-V_{1,\mu^k,\dagger}^{\mathcal{E}}(s_1)\right) && (18)\\
&=\mathbb{E}\left(\sum_{k=1}^K V_1^{\mathcal{E},*}(s_1)-V_{1,\mu^k,\nu^k}^{\mathcal{E}}(s_1)\right)+\mathbb{E}\left(\sum_{k=1}^K V_{1,\mu^k,\nu^k}^{\mathcal{E}}(s_1)-V_{1,\mu^k,\dagger}^{\mathcal{E}}(s_1)\right). && (19)
\end{aligned}$$

The first term in (19) represents the expected difference between the Nash value $V_1^{\mathcal{E},*}(s_1)$ and the value induced by the policy (μ^k,ν^k). The second term represents the Bayesian regret induced by the min-player's policy ν^k with respect to the fixed max-player's policy μ^k.

For the first term in (19), we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=1}^K V_1^{\mathcal{E},*}(s_1)-V_{1,\mu^k,\nu^k}^{\mathcal{E}}(s_1)\right)&\le\mathbb{E}\left(\sum_{k=1}^K V_{1,\mu^*(\mathcal{E}),\nu^k}^{\mathcal{E}}(s_1)-V_{1,\mu^k,\nu^k}^{\mathcal{E}}(s_1)\right) && (20)\\
&=\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}\left(V_{1,\mu^*(\mathcal{E}),\nu^k}^{\mathcal{E}}(s_1)-V_{1,\mu^k,\nu^k}^{\mathcal{E}}(s_1)\right)\right] && (21)\\
&\le\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\sqrt{\frac{\left(\mathbb{E}_k\left(V_{1,\mu^*(\mathcal{E}),\nu^k}^{\mathcal{E}}(s_1)-V_{1,\mu^k,\nu^k}^{\mathcal{E}}(s_1)\right)\right)^2}{\mathbb{I}_k^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k)}}\sqrt{\mathbb{I}_k^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k)}\right] && (22)\\
&\le\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\frac{\left(\mathbb{E}_k\left(V_{1,\mu^*(\mathcal{E}),\nu^k}^{\mathcal{E}}(s_1)-V_{1,\mu^k,\nu^k}^{\mathcal{E}}(s_1)\right)\right)^2}{\mathbb{I}_k^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k)}\right]}\cdot\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k)\right]} && (23)\\
&=\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\Gamma_k(\mu^k,\nu^k,\mathcal{E})\right]}\cdot\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k)\right]}, && (24)
\end{aligned}$$

where the expectation $\mathbb{E}_k(\cdot)$ is over the randomness of the environment $\mathcal{E}$ conditioned on the past trajectory $\mathcal{D}_k$, thus $\mathbb{E}_k(\cdot)$ is equivalent to $\mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}(\cdot)$ and these two notations may be used interchangeably. Eqn. (20) follows from the property of the NE policy introduced in (8). Eqn. (23) follows from the Cauchy-Schwarz inequality, and Eqn. (24) is due to the definition of information ratio.

Since the max-player chooses μ_{IDS}^k and the min-player chooses ν_{IDS}^k at episode k, we have

$$\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\Gamma_k(\mu_{\text{IDS}}^k,\nu_{\text{IDS}}^k,\mathcal{E})\right]}\le\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu\in\Pi_B}\Gamma_k(\mu_{\text{IDS}}^k,\nu,\mathcal{E})\right]}\le\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu\in\Pi_B}\Gamma_k(\mu_{\text{TS}}^k,\nu,\mathcal{E})\right]}, \tag{25}$$

where the last inequality introduces the TS policy and it holds since μ_{IDS}^k minimizes $\max_\nu\Gamma_k(\mu,\nu,\mathcal{E})$ by definition. Below, we present a result that the information ratio of the TS policy μ_{TS}^k is bounded for any $\nu\in\Pi_B$ and any distribution $\mathcal{E}$.

Lemma 1. *For any $\nu\in\Pi_B$ and any distribution $\mathcal{E}$, we have*

$$\Gamma_k(\mu_{TS}^k,\nu,\mathcal{E})\le 4H^3SAB.$$

Proof of Lemma 1. The detailed proof is deferred to Appendix B.1. □

Applying Lemma 1 to Eqn. (25), we obtain

$$\sqrt{\sum_{k=1}^K\mathbb{E}_{\mathcal{D}_k}\left[\Gamma_k(\mu_{\text{IDS}}^k,\nu_{\text{IDS}}^k,\mathcal{E})\right]}\le\sqrt{4KH^3SAB}. \tag{26}$$

For the second term of Eqn. (24), by standard information inequalities, we have

$$\sum_{k=1}^{K}\mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k)\right]=\sum_{k=1}^{K}\mathbb{I}^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k|\mathcal{D}_k) \tag{27}$$

$$=\sum_{k=1}^{K}\mathbb{I}^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}_{H+1}^k|\mathcal{T}_{H+1}^1,\dots,\mathcal{T}_{H+1}^{k-1}) \tag{28}$$

$$=\mathbb{I}^{\mu,\nu}(\mathcal{E};\mathcal{T}_{H+1}^1,\dots,\mathcal{T}_{H+1}^k) \tag{29}$$

$$=\mathbb{I}^{\mu,\nu}(\mathcal{E};\mathcal{D}_{K+1}), \tag{30}$$

where (27) follows from the definition of conditional mutual information, and (29) follows from the chain rule of mutual information.

Lemma 2. *For any policy (μ,ν), the mutual information between the environment $\mathcal{E}$ and the whole trajectory $\mathcal{D}_{K+1}$ induced by (μ,ν) is upper-bounded as $\mathbb{I}^{\mu,\nu}(\mathcal{E};\mathcal{D}_{K+1})\le 2S^2ABH\log(SKH)$.*

Proof of Lemma 2. The proof is adapted from [12, Lemma 3.3] with appropriate modifications for zero-sum MGs. □

Since Lemma 2 holds for any policy (μ,ν), it also applies to the executed policy $(\mu_{\text{IDS}},\nu_{\text{IDS}})=(\{\mu_{\text{IDS}}^k\}_{k\in[K]},\{\nu_{\text{IDS}}^k\}_{k\in[K]})$. Therefore, combining Eqns. (20)-(24), (26)-(29), and Lemma 2, as well as substituting (μ,ν) to $(\mu_{\text{IDS}},\nu_{\text{IDS}})$, we have

$$\mathbb{E}\left(\sum_{k=1}^{K}V_1^{\mathcal{E},*}(s_1)-V_{1,\mu_{\text{IDS}}^k,\nu_{\text{IDS}}^k}^{\mathcal{E}}(s_1)\right)\le\sqrt{4KH^3SAB}\cdot\sqrt{2S^2ABH\log(SKH)}\le 4S^{3/2}ABH^2\sqrt{K\log(SKH)}. \tag{31}$$

The second term in (19), which represents the Bayesian regret induced by the min-player's policy ν_{IDS}^k with respect to the fixed max-player's policy μ_{IDS}^k, can be handled in a similar manner. In Appendix B.2, we show that

$$\mathbb{E}\left(\sum_{k=1}^{K}V_{1,\mu_{\text{IDS}}^k,\nu_{\text{IDS}}^k}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{IDS}}^k,\dagger}^{\mathcal{E}}(s_1)\right)\le 4S^{3/2}ABH^2\sqrt{K\log(SKH)}. \tag{32}$$

Substituting Eqns. (31)-(32) into Eqn. (19), we complete the proof of Theorem 1.

5. The Reg-MAIDS algorithm

Although the basic algorithm MAIDS has been proven to be sample-efficient, it has a high computational complexity due to the requirement of optimizing over the policy spaces of two players (Π_A and Π_B) and calculating the mutual information term. To mitigate the computational load of MAIDS, we propose a more (computationally) efficient IDS algorithm—called Reg-MAIDS—that has the same Bayesian regret bound as MAIDS in zero-sum MGs while enjoying less computational complexity. Again, the learning target χ is selected as the environment $\mathcal{E}$. This approach is inspired by the prior work [12] on single-agent RL.

5.1. Algorithm descriptions

At the beginning of episode k, the max-player first calculates the posterior distribution of the environment $\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)$, and then chooses the policy

$$\mu_{\text{R-IDS}}^k=\arg\max_{\mu\in\Pi_A}\min_{\nu\in\Pi_B}\left\{\mathbb{E}_k[V_{1,\mu,\nu}^{\mathcal{E}}(s_1)]+\lambda\mathbb{I}_k^{\mu,\nu}(\mathcal{E};\mathcal{T}_{H+1}^k)\right\},$$

where $\lambda>0$ is a parameter that controls the relative weights of the expected value and the regularization term (i.e., the amount of information learned by the policy (μ,ν)). Compared to the joint information ratio term of MAIDS in (13), a key difference is that $\mu_{\text{R-IDS}}^k$ does not require the calculation of the value $V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)$. As we will see later in this section, this modification leads to a significant reduction in the complexity of searching for the optimal solution.

The min-player then chooses her policy based on the knowledge of $\mu_{\text{R-IDS}}^k$ and the posterior distribution $\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)$:

$$\nu_{\text{R-IDS}}^k=\arg\min_{\nu\in\Pi_B}\left\{\mathbb{E}_k[V_{1,\mu_{\text{R-IDS}}^k,\nu}^{\mathcal{E}}(s_1)]-\widetilde{\lambda}\mathbb{I}_k^{\mu_{\text{R-IDS}}^k,\nu}(\mathcal{E};\mathcal{T}_{H+1}^k)\right\},$$

where $\widetilde{\lambda}>0$ is a different parameter. Below, we first present a regret bound for Reg-MAIDS, and then introduce computationally efficient implementation methods (with pseudocode available in Algorithm 2).

5.2. Regret bounds

Theorem 3 below shows that Reg-MAIDS achieves the same Bayesian regret bound as MAIDS when setting λ and $\widetilde{\lambda}$ appropriately.

Algorithm 2 REG-MAIDS for two-player zero-sum Markov Games.

Input: Prior distribution ρ, parameters $\lambda > 0, \tilde{\lambda} > 0$, posterior sampling oracle (via epistemic neural networks)
for episode $k = 1$ **to** K **do**
  Both players call the oracle multiple times to approximate the mean environments $\bar{e}_k$ and $\bar{e}'_k$ (defined in Sec. 5.3).
  The max-player applies an efficient method (e.g., dynamic programming) to calculate the NE policy

$$\mu^k_{\text{R-IDS}} = \arg\max_{\mu \in \Pi_A} \min_{\nu \in \Pi_B} V^{\bar{e}_k}_{1,\mu,\nu}(s_1)$$

  The min-player applies an efficient method (e.g., dynamic programming) to calculate the best response to $\mu^k_{\text{R-IDS}}$:

$$\nu^k_{\text{R-IDS}} = \arg\min_{\nu \in \Pi_B} V^{\bar{e}'_k}_{1,\mu^k_{\text{R-IDS}},\nu}(s_1).$$

end for

Theorem 3. *Suppose the max-player's policy is $\mu_{R\text{-}IDS} = \{\mu^k_{R\text{-}IDS}\}_{k \in [K]}$ and the min-player's policy is $\nu_{R\text{-}IDS} = \{\nu^k_{R\text{-}IDS}\}_{k \in [K]}$, with $\lambda = \widetilde{\lambda} = \sqrt{2KH^2/S \log(SKH)}$. For any prior ρ, the Bayesian regret of $\mu_{R\text{-}IDS}$ satisfies*

$$\mathsf{BR}_K(\mu_{R\text{-}IDS}) \leq 8\sqrt{H^4 S^3 A^2 B^2 K \log(SKH)}. \tag{33}$$

Proof of Theorem 3. See Appendix D for the detailed proof. □

Similar to Theorem 2, when Assumption 1 holds, one can derive a refined regret bound of the form $\mathsf{BR}_K(\mu_{\text{R-IDS}}) \leq 8\sqrt{H^4 S^2 ABK \log(SKH)} + K_0 H$, which is valid for all K exceeding a sufficiently large constant K_0.

5.3. Equivalent formulas of $\mu^k_{R\text{-}IDS}$ and $\nu^k_{R\text{-}IDS}$

Given the past trajectory $\mathcal{D}_k$, we define a *mean environment* as $\bar{e}_k = (H, \mathcal{S}, \mathcal{A}, \mathcal{B}, \{P^{\bar{e}_k}_h\}_{h=1}^H, \{r^{\bar{e}_k}_h\}_{h=1}^H)$, where the transition kernel

$$P^{\bar{e}_k}_h(\cdot|s,a,b) \triangleq \mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}[P^{\mathcal{E}}_h(\cdot|s,a,b)], \quad \forall (s,a,b,h) \in (\mathcal{S},\mathcal{A},\mathcal{B},[H])$$

and the reward function

$$r^{\bar{e}_k}_h(s,a,b) \triangleq r_h(s,a,b) + \lambda \mathbb{E}_k\left[\mathbb{D}_{\text{KL}}(P^{\mathcal{E}}_h(\cdot|s,a,b) \| P^{\bar{e}_k}_h(\cdot|s,a,b))\right], \quad \forall (s,a,b,h) \in (\mathcal{S},\mathcal{A},\mathcal{B},[H]).$$

The definition of $P^{\bar{e}_k}_h(\cdot|s,a,b)$ is also called *Bayesian model averaging* [54] in statistics. For the max-player, Lemma 3 below shows that the objective function $\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu,\nu}(s_1)\right] + \lambda \mathbb{I}^{\mu,\nu}_k(\mathcal{E}; \mathcal{T}^k_{H+1})$ in the formula of $\mu^k_{\text{R-IDS}}$ is equal to the value $V^{\bar{e}_k}_{1,\mu,\nu}(s_1)$ in the mean environment $\bar{e}'_k$.

Lemma 3. *For any policy (μ,ν), we have*

$$\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu,\nu}(s_1)\right] + \lambda \mathbb{I}^{\mu,\nu}_k(\mathcal{E}; \mathcal{T}^k_{H+1}) = V^{\bar{e}_k}_{1,\mu,\nu}(s_1).$$

Proof of Lemma 3. See Appendix F. □

Given Lemma 3, one can rewrite the formula of $\mu^k_{\text{R-IDS}}$ as

$$\mu^k_{\text{R-IDS}} = \arg\max_{\mu \in \Pi_A} \min_{\nu \in \Pi_B} V^{\bar{e}_k}_{1,\mu,\nu}(s_1), \tag{34}$$

which is equivalent to finding a max-player's NE policy μ^* in the environment $\bar{e}_k$. For the min-player, we define $\bar{e}'_k = (H, \mathcal{S}, \mathcal{A}, \mathcal{B}, \{P^{\bar{e}'_k}_h\}_{h=1}^H, \{r^{\bar{e}'_k}_h\}_{h=1}^H)$ as another mean environment, where $P^{\bar{e}'_k}_h = P^{\bar{e}_k}_h$ and $r^{\bar{e}'_k}_h(s,a,b) \triangleq r_h(s,a,b) - \widetilde{\lambda} \mathbb{E}_k\left[\mathbb{D}_{\text{KL}}(P^{\mathcal{E}}_h(\cdot|s,a,b) \| P^{\bar{e}_k}_h(\cdot|s,a,b))\right]$. Similar to Lemma 3, one can show that for any μ,

$$\mathbb{E}_k[V^{\mathcal{E}}_{1,\mu,\nu}(s_1)] - \widetilde{\lambda} \mathbb{I}^{\mu,\nu}_k(\mathcal{E}; \mathcal{T}^k_{H+1}) = V^{\bar{e}'_k}_{1,\mu,\nu}(s_1).$$

Thus, the formula of $\nu^k_{\text{R-IDS}}$ can be rewritten as

$$\nu^k_{\text{R-IDS}} = \arg\min_{\nu \in \Pi_B} V^{\bar{e}'_k}_{1,\mu^k_{\text{R-IDS}},\nu}(s_1), \tag{35}$$

which is equivalent to finding the best response to the max-player's policy $\mu^k_{\text{R-IDS}}$ in the environment $\bar{e}'_k$.

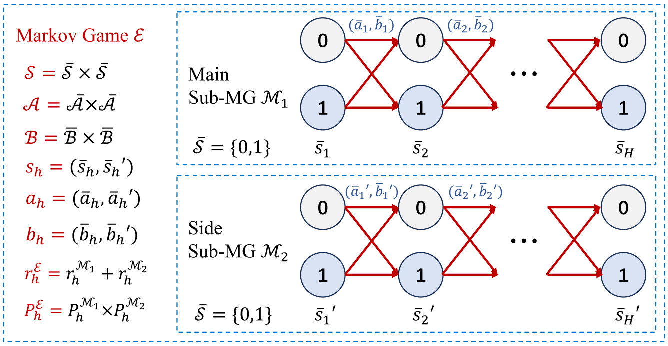

Fig. 1. Illustration of the MG in Example 1.

5.4. *Implementing* Reg-MAIDS *efficiently*

Following the prior work [12] in RL settings, we assume the existence of a *posterior sampling oracle*, which, upon receiving a call, outputs a sample of the environment $\mathcal{E}$ from the posterior distribution $\mathbb{P}(\cdot|\mathcal{D}_k)$. Multiple calls to the oracle lead to multiple *independent* samples. As discussed in [12], the posterior sampling oracle can be obtained, either exactly or approximately, using epistemic neural networks [55].

Given the posterior sampling oracle, the max-player can accurately approximate the mean environment $\bar{e}_k$ with transition kernel $\{P_h^{\bar{e}_k}\}_{h=1}^H$ and reward function $\{r_h^{\bar{e}_k}\}_{h=1}^H$ using Monte Carlo sampling. Next, based on the equivalent formula of $\mu^k_{\text{R-IDS}}$ in (34), the max-player can find a NE policy μ^* in the environment $\bar{e}'_k$ efficiently by using dynamic programming methods like value iteration or policy iteration. Similarly, the min-player can first approximate the mean environment $\bar{e}'_k$ using Monte Carlo sampling, and then finds the best response to the max-player's policy $\mu^k_{\text{R-IDS}}$ in the environment $\bar{e}'_k$ using dynamic programming methods like value iteration or policy iteration. The pseudocode of Reg-MAIDS is provided in Algorithm 2.

6. Learning compressed environments

An appealing feature of IDS is that it provides freedom and flexibility to select the learning target χ. In situations where the environment is overwhelmingly complex and surpasses agents' computation capacity, it can be beneficial for agents to focus only on essential parts of the environment, while disregarding less significant details that have less impact on decision-making (as illustrated in Example 1 below). This is analogous to the concept of lossy compression in information theory.

Example 1. Suppose an MG (environment) $\mathcal{E}$ can be decomposed into the *product* of two sub-MGs: the main sub-MG $\mathcal{M}_1$ and side sub-MG $\mathcal{M}_2$, each with state space $\bar{\mathcal{S}}$ and action spaces $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}$. An illustration is provided in Fig. 1. Assume the reward of side sub-MG $\mathcal{M}_2$ at each step is upper-bounded by δ. Intuitively, as $\delta \to 0$, the impact of $\mathcal{M}_2$ on the overall decision-making becomes increasingly insignificant. Consequently, agents can prioritize learning about $\mathcal{M}_1$ (while ignoring $\mathcal{M}_2$) in order to make good decisions.[6]

Following this intuition, we introduce new IDS-based algorithms where the learning target χ is not the environment $\mathcal{E}$ itself but a compressed environment.[7] We first propose a distortion measure to evaluate the discrepancy between any pair of environments, and then put forth two principles for constructing compressed environments.

6.1. *Distortion measure*

Inspired by *lossy compression* (a.k.a. rate-distortion theory) in information theory, we propose to compress the environment $\mathcal{E}$ to a compressed environment $\tilde{\mathcal{E}}$ that is "close" to $\mathcal{E}$ under a carefully defined distortion measure. Recall that Π_A and Π_B are the sets of max-player's and min-player's Markov policies respectively, and let $\Phi_A \subseteq \Pi_A$ and $\Phi_B \subseteq \Pi_B$ be any subsets of their policies. For any environments $e, e' \in \Theta$, we define

$$d_{\Phi_A,\Phi_B}(e,e') \triangleq \sup_{(\mu,\nu)\in\Phi_A\times\Phi_B} \left| V^e_{1,\mu,\nu}(s_1) - V^{e'}_{1,\mu,\nu}(s_1) \right| \tag{36}$$

6 Example 1 is conceptually similar to the Multi-Resolution MDP [22] introduced for single-agent settings. The Multi-Resolution MDP is more complex compared to our Example 1, as it contains N sub-environments where the reward of the n-th sub-environment is bounded in the interval $[0, 1/n]$. Similar to Example 1, the agent can effectively ignore the sub-environments whose contribution to the overall reward are negligible. However, we prefer our example as it allows us to easily quantify the performance gain of our proposed Compressed-MAIDS algorithm in this specific setting (see Section 6.5 for more details).

7 The *surrogate environment* introduced in [12,46] is conceptually similar to the compressed environment. We call it compressed environment because it is analogous to lossy compression problem in information theory.

as a distortion measure between e and e' through their value under the worst-case policy $(\mu,\nu)\in\Phi_A\times\Phi_B$. A comparison to the distortion measure proposed in [22] is provided in Remark 4, Appendix E.

Remark 2. A sufficient condition for the distortion $d_{\Phi_A,\Phi_B}(e,e')$ to be small is that the transition kernels $P_h^e(\cdot|s,a,b)$ and $P_h^{e'}(\cdot|s,a,b)$ are close enough for *all* (s,a,b,h). However, this condition may not be necessary. For instance, consider two environments e and e' in the context of Example 1. Suppose e and e' share the same main sub-MG $\mathcal{M}_1$ but differ in their side sub-MGs. In this case, even though $P_h^e(\cdot|s,a,b)$ and $P_h^{e'}(\cdot|s,a,b)$ are not close enough for some (s,a,b,h), the distortion still satisfies $d_{\Phi_A,\Phi_B}(e,e')\le 2H\delta$ for any $\Phi_A\subseteq\Pi_A$ and $\Phi_B\subseteq\Pi_B$.

6.2. Construction of compressed environments

Given the distribution of the environment $\mathcal{E}$ and the distortion measure d_{Φ_A,Φ_B}, we introduce two methods for constructing the compressed environment $\tilde{\mathcal{E}}$ (which is a random variable).

1. (Soft-compression) Following standard approaches in rate-distortion theory, we require the compressed environment $\tilde{\mathcal{E}}$ to be correlated with $\mathcal{E}$ in such a way that

$$\mathbb{E}[d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}})]\le\epsilon \tag{37}$$

for some tolerance level $\epsilon>0$. In particular, a popular choice of $\tilde{\mathcal{E}}$ is the one that minimizes the mutual information with $\mathcal{E}$ subject to (37), i.e.,[8]

$$\tilde{\mathcal{E}}=\underset{\tilde{\mathcal{E}}':\mathbb{E}[d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}')]\le\epsilon}{\arg\min}\ \mathbb{I}(\mathcal{E};\tilde{\mathcal{E}}'). \tag{38}$$

The choice in (38) is adopted by [22] for their value-equivalent sampling algorithm in RL, and is also the standard choice for many lossy compression problems in information theory [19].

2. (Hard-compression) Another way to define the compressed environment $\tilde{\mathcal{E}}$ is to require the distortion between $\mathcal{E}$ and $\tilde{\mathcal{E}}$ to be small almost everywhere, i.e.,

$$\mathbb{P}(d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}})>\epsilon)=0. \tag{39}$$

This is more stringent compared to (37), since (37) only requires the distortion between the pair of random variables $(\mathcal{E},\tilde{\mathcal{E}})$ to be small in the *average* sense.[9] To achieve (39), a feasible approach is to partition the entire environment space Θ into C_ϵ disjoint subspaces $\{\Theta_c\}_{c=1}^{C_\epsilon}$, such that any two environments e and e' in the same subspace satisfy $d_{\Phi_A,\Phi_B}(e,e')\le\epsilon$. Then, we arbitrarily select an element $e_c\in\Theta_c$ as the reference of Θ_c, and construct $\tilde{\mathcal{E}}$ such that

$$\tilde{\mathcal{E}}=e_c \ \text{ if and only if } \ \mathcal{E}\in\Theta_c. \tag{40}$$

Below, we provide a concrete construction of the compressed environment $\tilde{\mathcal{E}}$ that satisfies (39) for any $\Phi_A\subseteq\Pi_A$ and $\Phi_B\subseteq\Pi_B$, based on partitioning the transition kernels. First, we introduce the concept of *covering number*.

Definition 1 *(covering number)*. For any $\delta\ge 0$, we say the set $C(\delta)\subseteq\Delta(\mathcal{S})$ is a δ-covering of the probability simplex $\Delta(\mathcal{S})$ w.r.t. the total variation distance $\mathbb{D}_{TV}$ if $\forall P\in\Delta(\mathcal{S})$, $\exists P'\in C(\delta)$ such that $\mathbb{D}_{TV}(P,P')\le\delta$. The δ-covering number of $\Delta(\mathcal{S})$ is then denoted as

$$\kappa(\delta)\triangleq\min\{n:\exists \text{ a } \delta\text{-covering of } \Delta(\mathcal{S}) \text{ with cardinality } n\}.$$

Next, we describe the construction of the subspaces $\{\Theta_c\}_{c=1}^{C_\epsilon}$ (with $C_\epsilon=\kappa(\epsilon/2H^2)^{SABH}$) as well as the construction of the compressed environment $\tilde{\mathcal{E}}$.

1. We partition the environment space Θ into $C_\epsilon\triangleq\kappa(\epsilon/2H^2)^{SABH}$ subspaces $\{\Theta_c\}_{c=1}^{C_\epsilon}$, where $\kappa(\epsilon/2H^2)$ is the covering number. For each $(s,a,b,h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$, we divide the support of $P_h(\cdot|s,a,b)$, which is $\Delta(\mathcal{S})$, into $\kappa(\epsilon/2H^2)$ balls of radius $\epsilon/2H^2$ (w.r.t. the total variation distance). Each ball is centered at one element of the set $C(\epsilon/2H^2)$, where $C(\epsilon/2H^2)$ is an $(\epsilon/2H^2)$-covering of $\Delta(\mathcal{S})$. Thus, for any two distributions P and $\tilde{P}$ in the same ball, by the triangle inequality, it is clear that

$$\mathbb{D}_{\mathrm{TV}}(P,\tilde{P})\le\mathbb{D}_{\mathrm{TV}}(P,P')+\mathbb{D}_{\mathrm{TV}}(P',\tilde{P})\le\epsilon/H^2$$

where P' is the ball center. Next, we "trim" these balls in such a way that if a distribution $P\in\Delta(\mathcal{S})$ appears in two or more balls, we remove it from all but one ball. This trimming operation ensures that the partitions of $\Delta(\mathcal{S})$ are disjoint. Since we need to

[8] In theory, one can efficiently compute the optimal $\tilde{\mathcal{E}}$ in (38) using the Blahut-Arimoto algorithm [56,57]. However, in practice, challenges arise when dealing with continuous random variables, and we refer the readers to [22] for a discussion of effective empirical methods.

[9] This is also the reason why we name the two principles "soft-compression" and "hard-compression" respectively.

Algorithm 3 COMPRESSED-MAIDS for two-player zero-sum Markov Games.

1: **Input:** Prior distribution ρ
2: **for** episode $k = 1$ **to** K **do**
3: Both players calculate the posterior of environment $\mathcal{E}$ based on prior ρ and trajectory $\mathcal{D}_k$ (as per (15))
4: Both players construct a compressed environment $\tilde{\mathcal{E}}$ (correlated with $\mathcal{E}$) via the rule in (38) or (40).
5: The max-player calculates $\tilde{\Gamma}_k(\mu, \nu, \tilde{\mathcal{E}})$ for each μ, ν, and chooses

$$\mu^k_{\text{C-IDS}} = \arg\min_{\mu\in\Pi_A} \max_{\nu\in\Pi_B} \tilde{\Gamma}_k(\mu, \nu, \tilde{\mathcal{E}}).$$

6: The min-player calculates $\tilde{\Lambda}_k^{\mu^k_{\text{C-IDS}}}(\nu, \mathcal{E})$ for each ν, and chooses

$$\nu^k_{\text{C-IDS}} = \arg\min_{\nu\in\Pi_B} \tilde{\Lambda}_k^{\mu^k_{\text{C-IDS}}}(\nu, \tilde{\mathcal{E}})$$

7: **end for**

divide the support of $P_h(\cdot|s,a,b)$ for all $(s,a,b,h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times[H]$, we have $C_\epsilon = \kappa(\epsilon/2H^2)^{SABH}$. It is clear that the subspaces $\{\Theta_c\}_{c=1}^{C_\epsilon}$ are disjoint and their union constitutes the space Θ.

2. For each subspace Θ_c, we arbitrarily select an element $e_c \in \Theta_c$ as the *reference environment* of Θ_c. We then construct the compressed environment $\tilde{\mathcal{E}}$ in such a way that

$$\tilde{\mathcal{E}} = e_c \quad \text{if and only if} \quad \mathcal{E} \in \Theta_c. \tag{41}$$

Lemma 7 in Appendix E.2 shows that the above construction of $\tilde{\mathcal{E}}$ satisfies the hard-compression constraint in (39).

6.3. *The* COMPRESSED-MAIDS *algorithm*

We now introduce the algorithm COMPRESSED-MAIDS, where the pseudocode is provided in Algorithm 3. At the beginning of episode k, the max-player first calculates the posterior distribution of the environment $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$, and then chooses a compressed environment $\tilde{\mathcal{E}}$ via the above construction rules (e.g., following (38) or (40)). Next, the max-player chooses the policy

$$\mu^k_{\text{C-IDS}} = \arg\min_{\mu\in\Pi_A} \max_{\nu\in\Pi_B} \tilde{\Gamma}_k(\mu, \nu, \tilde{\mathcal{E}}),$$
$$\text{where } \tilde{\Gamma}_k(\mu, \nu, \tilde{\mathcal{E}}) \triangleq \frac{(\mathbb{E}_k[V^{\tilde{\mathcal{E}}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu,\nu}(s_1)])^2}{\mathbb{I}_k^{\mu,\nu}(\tilde{\mathcal{E}}; \mathcal{T}^k_{H+1})}.$$

The min-player calculates the distributions of $\mathcal{E}$ and $\tilde{\mathcal{E}}$ in the same way as the max-player, and chooses her policy as:

$$\nu^k_{\text{C-IDS}} = \arg\min_{\nu\in\Pi_B} \tilde{\Lambda}_k^{\mu^k_{\text{C-IDS}}}(\nu, \tilde{\mathcal{E}}),$$
$$\text{where } \tilde{\Lambda}^{\mu}_k(\nu, \tilde{\mathcal{E}}) \triangleq \frac{(\mathbb{E}_k[V^{\tilde{\mathcal{E}}}_{1,\mu,\nu}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu,\nu^\dagger_{\mathcal{E}}(\mu)}(s_1)])^2}{\mathbb{I}_k^{\mu,\nu}(\tilde{\mathcal{E}}; \mathcal{T}^k_{H+1})}.$$

Here, $\nu^\dagger_{\mathcal{E}}(\mu)$ denotes the best response to the max-player's policy μ in the environment $\mathcal{E}$.

6.4. *Regret bounds*

Next, we present Bayesian regret bounds for COMPRESSED-MAIDS when the compressed environment $\tilde{\mathcal{E}}$ is constructed through hard-compression as in (40), while regret bounds in the context of soft-compression will be considered in future works. Below, we express $\tilde{\mathcal{E}}$ as $\tilde{\mathcal{E}}_\epsilon$ to clearly indicate its correlation with the parameter ϵ.

Theorem 4. *Let the max-player's policy is $\mu_{\text{C-IDS}} = \{\mu^k_{\text{C-IDS}}\}_{k=1}^K$ and the min-player's policy is $\nu_{\text{C-IDS}} = \{\nu^k_{\text{C-IDS}}\}_{k=1}^K$, where the compressed environment $\tilde{\mathcal{E}}_\epsilon$ at each episode $k \in [K]$ is constructed according to (40) with $\epsilon > 0$. Then for any prior ρ, the Bayesian regret of $\mu_{\text{C-IDS}}$ satisfies*

$$\mathsf{BR}_K(\mu_{\text{C-IDS}}) \le 4\sqrt{KH^3SAB\cdot\mathbb{I}(\tilde{\mathcal{E}}_\epsilon; \mathcal{D}_{K+1})} + 4K\epsilon,$$

where $\mathbb{I}(\tilde{\mathcal{E}}_\epsilon; \mathcal{D}_{K+1})$ is the mutual information w.r.t. executing the policy $(\mu_{\text{C-IDS}}, \nu_{\text{C-IDS}})$.

The proof of Theorem 4 is provided in Appendix E. Moreover, since one has the freedom to optimize ϵ, we have

$$\begin{aligned}\mathsf{BR}_K(\mu_{\text{C-IDS}}) &\le \inf_{\epsilon\ge 0}\left\{4\sqrt{KH^3SAB\cdot\mathbb{I}(\tilde{\mathcal{E}}_\epsilon; \mathcal{D}_{K+1})} + 4K\epsilon\right\}\\ &\le 4\sqrt{KH^3SAB\cdot\mathbb{I}(\tilde{\mathcal{E}}_{1/K}; \mathcal{D}_{K+1})} + 4.\end{aligned}$$

6.5. Advantages of COMPRESSED-MAIDS

We first discuss the advantages of COMPRESSED-MAIDS in the context of our illustrative example, Example 1, when the parameter $\delta = 1/2HK$. Considering the special structure of the MG in this example, an intuitive approach to compress the environment $\mathcal{E}$ is to retain the main sub-MG $\mathcal{M}_1$ while discard the (negligible) side sub-MG $\mathcal{M}_2$. Following the hard-compression principle, we first construct subspaces such that each subspace comprises all environments that have a same main sub-MG $\mathcal{M}_1$, upon which we construct a compressed environment $\tilde{\mathcal{E}}$ based on (40).[10] One can check that any pair of environments e and e' in the same subspace satisfy $d_{\Phi_A,\Phi_B}(e,e') \le 2H\delta = 1/K$, thus $\tilde{\mathcal{E}}$ can be expressed as $\tilde{\mathcal{E}}_{1/K}$. Applying an upper bound on the mutual information (see Lemma 2, Appendix B), we have $\mathbb{I}(\tilde{\mathcal{E}}_{1/K}; \mathcal{D}_{K+1}) \le 2|\bar{\mathcal{S}}|^2|\bar{\mathcal{A}}||\bar{\mathcal{B}}|H\log(|\bar{\mathcal{S}}|KH)$ where $|\bar{\mathcal{S}}| = \sqrt{S}$, $|\bar{\mathcal{A}}| = \sqrt{A}$ and $|\bar{\mathcal{B}}| = \sqrt{B}$. Thus, the Bayesian regret of $\mu_{\text{C-IDS}}$ becomes $\tilde{O}(\sqrt{H^4S^2(AB)^{3/2}K})$, which reduces a factor of $\sqrt{S(AB)^{1/2}}$ compared to the regret of the basic algorithm MAIDS (Theorem 1). This suggests COMPRESSED-MAIDS is more efficient, as employing MAIDS would lead the agents to concurrently learn about both the main and side sub-MGs, an approach that is less efficient considering the negligibility of the side sub-MG.

It is worth noting that the construction of $\tilde{\mathcal{E}}$ in (40) is just one specific choice of the compressed environment; consequently, its regret may not be optimal. However, the IDS principle offers the flexibility in selecting methods for constructing the compressed environment, and different constructions may lead to different performances/regrets. In the single-agent RL setting, [12, Sec 4.1] introduces a novel approach that helps to reduce the Bayesian regret by a factor of $\sqrt{S}$; however, their approach cannot be easily applied to MGs due to the presence of unpredictable opponents. While MARL is more complicated, we still conjecture that it is possible to construct alternative compressed environments satisfying (37) or (39) that can lead to better regret bounds. This opens up an intriguing avenue for future research.

7. General-sum Markov games

In this section, we introduce the mathematical formulation of the multi-player general-sum MG, and then show that an extended version of REG-MAIDS (Section 5) is provably sample-efficient in this general setting. In the following, let $N \ge 2$ be the number of players. A summary of notations used in the general-sum MG is provided in Appendix A.2.

7.1. Model

The general-sum MG can be represented by a tuple $\mathcal{E} = (N, H, \mathcal{S}, \mathcal{A}, \{P_h\}_{h\in[H]}, \{r_h^{(i)}\}_{h\in[H],i\in[N]})$, where $\mathcal{A} = \otimes_{i=1}^N \mathcal{A}_i$ is the joint action spaces of the N players, $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, and $r_h^{(i)} : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the deterministic reward function for player i. Let $S \triangleq |\mathcal{S}|$, $A_i \triangleq |\mathcal{A}_i|$, and $A \triangleq |\mathcal{A}| = \prod_{i=1}^N A_i$. Similar to the zero-sum MG setting (Section 3), we assume $\{r_h^{(i)}\}_{h\in[H],i\in[N]}$ are known, and we consider a Bayesian setting where we have a prior distribution ρ on the environment $\mathcal{E}$ (i.e., on the transition kernels $\{P_h\}_{h\in[H]}$).

For player i, let $\Pi_i^{\text{pure}} \subseteq \{\mathcal{S} \to \Delta(\mathcal{A}_i)\}^H$ be a set of (Markov) *pure policies*. She is also allowed to take a *random policy* by adopting a random seed $\omega \in \Omega$. Specifically, a random policy $\pi^{(i)} = \{\pi_h^{(i)}\}_{h\in[H]}$ can be viewed as a mapping from Ω to Π_i^{pure}, i.e., $\pi^{(i)}(\omega) \in \Pi_i^{\text{pure}}$. When executing a random policy $\pi^{(i)}$, she first samples a random seed ω and then follows $\pi^{(i)}(\omega) = \{\pi_h^{(i)}(\omega)\}_{h\in[H]}$ in the H steps. Next, we consider the N players jointly, and introduce the concepts of *joint policy* and *product policy*. The joint policy $\pi = \{\pi^{(i)}(\omega_i)\}_{i\in[N]}$ consists of N random policies where random variables $(\omega_1, \dots, \omega_N)$ are correlated, while the product policy $\pi = \{\pi^{(i)}(\omega_i)\}_{i\in[N]}$ consists of N random policies where $(\omega_1, \dots, \omega_N)$ are independent. As such, product policies are special cases of joint policies.

For policy π and player i, the value function is denoted as

$$V_{h,\pi}^{(i),\mathcal{E}}(s) = \mathbb{E}_\pi^{\mathcal{E}}\left[\sum_{h'=h}^{H} r_{h'}^{(i)}(s_{h'}, a_{h'}) \Big| s_h = s\right], \ \forall s \in \mathcal{S}.$$

For any policy π, let $\pi^{(-i)}$ be the joint policy excluding player i, and let

$$\pi_{\mathcal{E}}^{(i),\dagger} \triangleq \mathop{\arg\max}_{\nu\in\Delta(\Pi_i^{\text{pure}})} V_{1,\nu\times\pi^{(-i)}}^{(i),\mathcal{E}}(s_1)$$

be the best response of player i when facing $\pi^{(-i)}$, where s_1 is the fixed initial state.

7.2. Learning objectives

For the general-sum MG, we first define the concept of (approximate) NE and coarse correlated equilibrium (CCE).

Definition 2 *(ε-NE and ε-CCE).* A policy π is an ε-NE of the environment $\mathcal{E}$ if it is a *product policy* and it satisfies

$$V_{1,(\pi_{\mathcal{E}}^{(i),\dagger},\pi^{(-i)})}^{(i),\mathcal{E}}(s_1) - V_{1,\pi}^{(i),\mathcal{E}}(s_1) \le \varepsilon, \ \forall i \in [N]. \tag{42}$$

We say a policy π is an ε-CCE if it is a *joint policy* and it satisfies (42).

[10] For the purpose of illustration, we allow the number of subspaces to be infinite here (as the number of $\mathcal{M}_1$ may be infinite).

Algorithm 4 GENERAL-MAIDS for multi-player general-sum Markov Games.

1: **Input:** Prior distribution ρ
2: **for** episode $k = 1$ **to** K **do**
3: Calculate the posterior of environment $\mathcal{E}$ based on prior ρ and trajectory $\mathcal{D}_k$:

$$\mathbb{P}(\mathcal{E}|\mathcal{D}_k) \propto \rho(\mathcal{E}) \prod_{i=1}^{k-1} \prod_{h=1}^{H} P_h^{\mathcal{E}}(s_{h+1}^i | s_h^i, a_h^i). \tag{44}$$

4: Calculate the mean environment $\hat{e}_k$ as per (45) and (46).
5: For each pure policy $\pi \in \otimes_{i=1}^N \Pi_i^{\text{pure}}$, calculate $V_{1,\pi}^{(i),\hat{e}_k}(s_1)$ for player $i \in [N]$.
6: View $\{V_{1,\pi}^{(i),\hat{e}_k}(s_1)\}_{i\in[N],\pi\in\otimes_{i=1}^N \Pi_i^{\text{pure}}}$ as payoff functions on a normal-form game, and computes the NE/CCE of the normal-form game.
7: **end for**

Suppose the N players interact with the environment for K episodes. For a *product policy* $\pi = \{\pi^k\}_{k\in[K]}$ and a realization of the environment $\mathcal{E} = e$, we define

$$\mathsf{Reg}_K^{\mathrm{NE}}(e, \pi) \triangleq \sum_{k=1}^{K} \sum_{i=1}^{N} V_{1,(\pi_e^{k,(i),\dagger}, \pi^{k,(-i)})}^{(i),e}(s_1) - V_{1,\pi^k}^{(i),e}(s_1)$$

as the cumulative regret in environment e, and define

$$\mathsf{BR}_K^{\mathrm{NE}}(\pi) \triangleq \mathbb{E}_{\mathcal{E}\sim\rho}\left(\mathsf{Reg}_K^{\mathrm{NE}}(\mathcal{E}, \pi)\right) \tag{43}$$

as the Bayesian regret, which is averaged over the prior distribution ρ of the environment. For a *joint policy* π, the definitions of regret $\mathsf{Reg}_K^{\mathrm{CCE}}(e, \pi)$ and Bayesian regret $\mathsf{BR}_K^{\mathrm{CCE}}(\pi)$ are exactly the same as $\mathsf{Reg}_K^{\mathrm{NE}}(e, \pi)$ and $\mathsf{BR}_K^{\mathrm{NE}}(\pi)$ (the only difference is that π is a joint policy).

7.3. Algorithm descriptions

We now extend REG-MAIDS to the general-sum MG, and we call this algorithm GENERAL-MAIDS. The pseudocode of GENERAL-MAIDS is provided in Algorithm 4. At the beginning of episode k, the posterior distribution of the environment $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$ can be calculated based on $\mathcal{D}_k$ and prior distribution ρ. Similar to Section 5, we construct a mean environment $\hat{e}_k$ such that the transition kernel and reward satisfy

$$P_h^{\hat{e}_k}(\cdot|s, a) \triangleq \mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}[P_h^{\mathcal{E}}(\cdot|s, a)], \tag{45}$$

$$r_h^{(i),\hat{e}_k}(s, a) \triangleq r_h^{(i)}(s, a) + \lambda\mathbb{E}_k[\mathbb{D}_{\mathrm{KL}}(P_h^{\mathcal{E}}(\cdot|s, a)\| P_h^{\hat{e}_k}(\cdot|s, a))],\ \forall i \in [N]. \tag{46}$$

The algorithm GENERAL-MAIDS proceeds in two steps:

1. For each pure policy $\pi \in \otimes_{i=1}^N \Pi_i^{\text{pure}}$, we calculate the value $V_{1,\pi}^{(i),\hat{e}_k}(s_1)$ in the mean environment $\hat{e}_k$ for all the players;
2. We view the MG on $\hat{e}_k$ as a *normal-form game* with the space of pure strategies being $\{\Pi_i^{\text{pure}}\}_{i\in[N]}$ and payoff functions being $\{V_{1,\pi}^{(i),\hat{e}_k}(s_1)\}_{i\in[N],\pi\in\otimes_{i=1}^N \Pi_i^{\text{pure}}}$. We then compute a NE strategy π^{NE} (resp. a CCE strategy π^{CCE}) for this normal-form game, and let the output policy for episode k in the MG be $\pi_{\text{G-IDS}}^{\mathrm{NE},k} = \pi^{\mathrm{NE}}$ (resp. $\pi_{\text{G-IDS}}^{\mathrm{CCE},k} = \pi^{\mathrm{CCE}}$).

Theorem 5 below provides Bayesian regret bounds for $\pi_{\text{G-IDS}}^{\mathrm{NE}} = \{\pi_{\text{G-IDS}}^{\mathrm{NE},k}\}_{k\in[K]}$ and $\pi_{\text{G-IDS}}^{\mathrm{CCE}} = \{\pi_{\text{G-IDS}}^{\mathrm{CCE},k}\}_{k\in[K]}$.

Theorem 5. *For any prior distribution ρ, the Bayesian regrets of $\pi_{\text{G-IDS}}^{NE}$ and $\pi_{\text{G-IDS}}^{CCE}$ respectively satisfy*

$$\mathsf{BR}_K^{NE}(\pi_{\text{G-IDS}}^{NE}) \leq 3N\sqrt{H^4 S^3 A^2 K \log(SKH)}, \quad \textit{and} \tag{47}$$

$$\mathsf{BR}_K^{CCE}(\pi_{\text{G-IDS}}^{CCE}) \leq 3N\sqrt{H^4 S^3 A^2 K \log(SKH)} \tag{48}$$

by choosing $\lambda = \sqrt{HK^2/S \log(SKH)}$.

The proof of Theorem 5 is provided in Appendix G.

Similar to the regret bound of REG-MAIDS for the two-player zero-sum MG, the regret bound in Theorem 5 for the multi-player general-sum MG is also valid for all possible prior distribution ρ of the environment $\mathcal{E}$, and the scaling $\tilde{O}(\sqrt{K})$ is also order-optimal w.r.t. the number of episodes K. It is worth noting that our regret bound for multi-player general-sum MGs recovers those for two-player zero-sum MGs (Theorems 1 and 3) when specializing to the two-player setting. Additionally, we point out that while this paper focuses on learning NE/CCE, it is also possible to learn a *correlated equilibrium* (CE) policy with theoretical guarantees via similar techniques.

Remark 3. In the context of multi-player general-sum Markov games, computing a Nash equilibrium (NE) or a Coarse Correlated Equilibrium (CCE) can be computationally challenging. Specifically, it is known that computing the NE is PPAD-complete. For CCE, however, the equilibrium can be approximated using mirror descent, such that the computational complexity is independent of the size of the pure policy space and instead depends only on the number of iterations in the approximation algorithm.

7.4. Learning compressed environments in the general-sum MG

Similar to the two-player zero-sum setting, one can also change the learning target from the entire environment $\mathcal{E}$ to a compressed environment $\tilde{\mathcal{E}}$ and design corresponding MARL algorithms in the multi-player general-sum MG. Specifically, we define a distortion measure $d : \Theta \times \Theta \to \mathbb{R}$ between any pair of environments (e, e') as

$$d(e,e') \triangleq \max_{i\in[N]} \sup_{\pi\in\Delta(\Pi_1^{pure})\times\cdots\times\Delta(\Pi_N^{pure})} \left| V_{1,\pi}^{(i),e}(s_1) - V_{1,\pi}^{(i),e'}(s_1) \right|, \tag{49}$$

and construct a compressed environment $\tilde{\mathcal{E}}$ that satisfies either (i) the soft-compression principle $\mathbb{E}[d(\mathcal{E},\tilde{\mathcal{E}})] \le \epsilon$, or (ii) the hard-compression principle $\mathbb{P}(d(\mathcal{E},\tilde{\mathcal{E}}) > \epsilon) = 0$. A concrete construction that satisfies both the soft-compression and hard-compression principles can be derived in a manner analogous to the approach used for the two-player zero-sum setting (see Appendix E.2).

We now introduce an algorithm that incorporates the compressed environment as the learning target. At the beginning of each episode k, both the environment $\mathcal{E} \sim \mathbb{P}(\cdot|\mathcal{D}_k)$ and the compressed version $\tilde{\mathcal{E}}$ are calculated based on $\mathcal{D}_k$ and the prior ρ. In contrast to the mean environment defined in (45)-(46), we define a modified mean environment $\check{e}_k$ such that the transition kernel and reward satisfy

$$P_h^{\check{e}_k}(\cdot|s,a) \triangleq \mathbb{E}_k[P_h^{\tilde{\mathcal{E}}}(\cdot|s,a)], \tag{50}$$

$$r_h^{(i),\check{e}_k}(s,a) \triangleq r_h^{(i)}(s,a) + \lambda \mathbb{E}_k[\mathbb{D}_{\mathrm{KL}}(P_h^{\tilde{\mathcal{E}}}(\cdot|s,a) \| P_h^{\check{e}_k}(\cdot|s,a))], \ \forall i \in [N]. \tag{51}$$

The remaining steps of the algorithm follow the same procedure as outlined in Sec. 7.3, with the only modification being the replacement of the calculated value from $V_{1,\pi}^{(i),\hat{e}_k}(s_1)$ to $V_{1,\pi}^{(i),\check{e}_k}(s_1)$. The resulting policies are denoted as $\pi_{\text{CG-IDS}}^{\text{NE}} = \{\pi_{\text{CG-IDS}}^{\text{NE},k}\}_k$ (for learning NE) and $\pi_{\text{CG-IDS}}^{\text{CCE}} = \{\pi_{\text{CG-IDS}}^{\text{CCE},k}\}_k$ (for learning CCE). Below, we present the corresponding theoretical guarantees, where we express $\tilde{\mathcal{E}}$ as $\tilde{\mathcal{E}}_\epsilon$ to clearly indicate its correlation with the parameter ϵ.

Theorem 6. *For any prior distribution ρ, the Bayesian regrets of $\pi_{\text{CG-IDS}}^{NE}$ and $\pi_{\text{CG-IDS}}^{CCE}$ respectively satisfy*

$$\mathsf{BR}_K^{NE}(\pi_{\text{CG-IDS}}^{NE}) \le 2N\sqrt{H^3SAK\cdot\mathbb{I}(\tilde{\mathcal{E}}_\epsilon;\mathcal{D}_{K+1})} + 2KN\epsilon, \quad \text{and} \tag{52}$$

$$\mathsf{BR}_K^{CCE}(\pi_{\text{CG-IDS}}^{CCE}) \le 2N\sqrt{H^3SAK\cdot\mathbb{I}(\tilde{\mathcal{E}}_\epsilon;\mathcal{D}_{K+1})} + 2KN\epsilon, \tag{53}$$

for any $\epsilon > 0$, where we choose $\lambda = \sqrt{SAH^3K/\mathbb{I}(\tilde{\mathcal{E}}_\epsilon;\mathcal{D}_{K+1})}$ in $\check{e}_k$, and $\mathbb{I}(\tilde{\mathcal{E}}_\epsilon;\mathcal{D}_{K+1})$ is the mutual information w.r.t. executing the policy $\pi_{\text{CG-IDS}}^{NE}$ or $\pi_{\text{CG-IDS}}^{CCE}$. Note that by further choosing $\epsilon = 1/K$, the RHS of (52) or (53) becomes $2N\sqrt{H^3SAK\cdot\mathbb{I}(\tilde{\mathcal{E}}_{1/K};\mathcal{D}_{K+1})} + 2N$.

A proof sketch of Theorem 6 is provided in Appendix H, which relies on a combination of the proof techniques for Theorems 4 and 5. The advantage of learning the compressed environment are more apparent when the general-sum MG exhibits a structure similar to that of Example 1, where certain portions of the environmental information are redundant.

8. Conclusion and future works

This work presents sample-efficient MARL algorithms based on the design principle of information-directed sampling (IDS). These algorithms utilize information-theoretic concepts to effectively navigate the exploration-exploitation tradeoff by balancing the acquired information about the environment (exploration) and the policy's sub-optimality (exploitation) when choosing policies. Theoretically, we prove that these IDS-based algorithms achieve order-optimal Bayesian regret with respect to the number of episodes in both two-player zero-sum MGs and multi-player general-sum MGs. This contribution enriches the set of sample-efficient algorithms available for tackling MARL problems. Moreover, we also provide a computationally efficient variant, Reg-MAIDS, that is not only theoretically sound but can also be implemented practically.

Finally, we put forth two promising directions for future research.

1. While the Bayesian regret bounds for our IDS-based algorithms are order-optimal with respect to the number of episodes K, they do not achieve order-optimality in terms of the cardinality of state space S, the cardinality of action spaces, and the episode length H. One would expect to close this gap by refining the algorithm design (e.g., choosing other compressed environments as the learning target in Compressed-MAIDS). Working in this direction would be a fruitful endeavor for better justifying the theoretical advantages of the IDS principle.
2. As an initial investigation of the IDS principle in MARL, this work mainly focuses on the *tabular* setting where the state space and the action space are discrete and finite. To address more practical scenarios where the state space is exceedingly large, it would be beneficial to develop new IDS-based algorithms for MG with *linear function approximation* or *general function approximation*,

enhancing their applicability in real-world scenarios. Moreover, we note that our IDS-based algorithms are self-play algorithms that follow the *Centralized Training with Decentralized Execution* paradigm. Exploring a fully decentralized version in future work could be an interesting direction.

CRediT authorship contribution statement

Qiaosheng Zhang: Writing – original draft, Methodology. **Chenjia Bai:** Methodology. **Shuyue Hu:** Methodology, Writing – original draft. **Zhen Wang:** Methodology, Writing – review & editing. **Xuelong Li:** Writing – review & editing, Methodology.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Qiaosheng Zhang, Zhen Wang reports financial support was provided by National Natural Science Foundation of China (No. U22B2036). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank Dr. Zhuoran Yang and Dr. Han Qi for their valuable discussions and insights, which have significantly improved the quality of this paper.

This work is supported by the National Science Fund for Distinguished Young Scholars (No. 62025602), the National Natural Science Foundation of China (No. U22B2036), the Tencent Foundation and XPLORER PRIZE, and Shanghai Artificial Intelligence Laboratory.

Appendix A. Summary of notations

A.1. *Notations for two-player zero-sum MGs*

$\Delta(\mathcal{X})$	Probability simplex over the set $\mathcal{X}$
$\mathcal{S}$	State space
$\mathcal{A}, A$	Action space of the max-player, cardinality of $\mathcal{A}$
$\mathcal{B}, B$	Action space of the min-player, cardinality of $\mathcal{B}$
H	Length of each episode
P_h	Transition kernel at step h ($P_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \Delta(\mathcal{S})$)
r_h	Reward function at step h ($r_h : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to [0,1]$)
$\mathcal{E}$	Environment (random variable)
e	Environment (realization)
K	Number of episodes
Θ	Parameter space of the transition kernels $\{P_h\}_{h=1}^H$
ρ	Prior distribution of the kernels $\{P_h\}_{h=1}^H$
$\mathcal{T}_h^k$	Trajectory at episode k up to step h
$\mathcal{D}_k$	Full trajectory up to the beginning of episode k
Π_A/Π_B	Set of all Markov policies of the max-player/min-player
$V_{h,\mu,\nu}^{\mathcal{E}}(s)$	Value of state s at step h (w.r.t. policy (μ,ν) and environment $\mathcal{E}$)
$\mu^\dagger(\nu)$	Best response to policy ν
$V_{h,\dagger,\nu}^{\mathcal{E}}(s)$	Value of state s w.r.t. ν and $\mu^\dagger(\nu)$ (i.e., $V_{h,\dagger,\nu}^{\mathcal{E}}(s) = V_{h,\mu^\dagger(\nu),\nu}^{\mathcal{E}}(s)$)
(μ^*,ν^*)	Nash equilibrium policy
$V_h^{\mathcal{E},*}(s)$	Nash value of state s at step h w.r.t. environment $\mathcal{E}$
χ	Learning target
$\mathbb{E}_k(\cdot)$	Expectation conditioned on the past trajectory (i.e., $\mathbb{E}_k(\cdot) = \mathbb{E}(\cdot\vert\mathcal{D}_k)$)
$\mathbb{I}_k(\cdot;\cdot)$	Mutual information conditioned on $\mathcal{D}_k$
$\Gamma(\mu,\nu,\chi)$	Information ratio
$\Lambda_k^\mu(\nu,\chi)$	Conditional information ratio w.r.t. policy μ
$\bar{e}_k, \bar{e}'_k$	Mean environments defined in Section 5
Φ_A/Φ_B	Subset of Markov policies of the max-player/min-player
$d_{\Phi_A,\Phi_B}(\cdot,\cdot)$	Distortion measure w.r.t. Φ_A and Φ_B
$\tilde{\mathcal{E}}$	Compressed environment
C_δ	δ-covering of probability simplex $\Delta(\mathcal{S})$
$\kappa(\delta)$	δ-covering number of probability simplex $\Delta(\mathcal{S})$

A.2. Notations for multi-player general-sum MGs

N	Number of players
$\mathcal{S}$	State space
$\mathcal{A}_i$	Action space of player i
$\mathcal{A}$	Joint action spaces of the N players
P_h	Transition kernel at step h ($P_h : \mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$)
$r_h^{(i)}$	Reward function for player i at step h ($r_h^{(i)} : \mathcal{S}\times\mathcal{A}\to[0,1]$)
Π_i^{pure}	Set of (Markov) *pure policies* of player i
ω	Random seed
$V_{h,\pi}^{(i),\mathcal{E}}(s)$	Value of state s for player i (at step h, w.r.t. policy π and environment $\mathcal{E}$)
$\pi^{(-i)}$	Policy π excluding player i
$\pi_{\mathcal{E}}^{(i),\dagger}$	Best response of player i when facing $\pi^{(-i)}$ in environment $\mathcal{E}$

Appendix B. Detailed proofs for MAIDS

B.1. Proof of Lemma 1

Recall that

$$\Gamma_k(\mu_{\text{TS}}^k,\nu,\mathcal{E})=\frac{\left(\mathbb{E}_k\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\mathcal{E}}(s_1)\right]\right)^2}{\mathbb{I}_k^{\mu_{\text{TS}}^k,\nu}(\mathcal{E};\mathcal{T}_{H+1}^k)}. \tag{B.1}$$

Let us focus on the numerator of the RHS of (B.1) from now on. It is worth noting that the term $\mathbb{E}_k[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\mathcal{E}}(s_1)]$ may not always be positive; thus, to provide an upper bound on the numerator of the RHS of (B.1), one need to upper-bound the absolute value of $\mathbb{E}_k[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\mathcal{E}}(s_1)]$.

Given the past trajectory $\mathcal{D}_k$, we define an environment $\tilde{e}_k=(H,\mathcal{S},\mathcal{A},\mathcal{B},\{P_h^{\tilde{e}_k}\}_{h=1}^H,\{r_h\}_{h=1}^H)$, where the transition kernel

$$P_h^{\tilde{e}_k}(\cdot|s,a,b)\triangleq\mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}[P_h^{\mathcal{E}}(\cdot|s,a,b)].$$

Note that $\tilde{e}_k$ is similar to the mean environments $\bar{e}_k$ and $\bar{e}_k'$ defined in Section 5, but differs from them in terms of the reward functions. Based on $\tilde{e}_k$, we now decompose the absolute value of $\mathbb{E}_k[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\mathcal{E}}(s_1)]$ as follows:

$$\begin{aligned}&\left|\mathbb{E}_k\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\mathcal{E}}(s_1)\right]\right|\\&\le\left|\mathbb{E}_k\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\tilde{e}_k}(s_1)\right]\right|+\left|\mathbb{E}_k\left[V_{1,\mu_{\text{TS}}^k,\nu}^{\tilde{e}_k}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\mathcal{E}}(s_1)\right]\right|.\end{aligned} \tag{B.2}$$

For the first term in (B.2), we have

$$\left|\mathbb{E}_k\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu_{\text{TS}}^k,\nu}^{\tilde{e}_k}(s_1)\right]\right| \tag{B.3}$$

$$=\left|\mathbb{E}_k\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)\right]-V_{1,\mu_{\text{TS}}^k,\nu}^{\tilde{e}_k}(s_1)\right| \tag{B.4}$$

$$=\left|\mathbb{E}_k\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)\right]-\mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\tilde{e}_k}(s_1)\right]\right| \tag{B.5}$$

$$=\left|\mathbb{E}_k\left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1)-V_{1,\mu^*(\mathcal{E}),\nu}^{\tilde{e}_k}(s_1)\right]\right|, \tag{B.6}$$

where Eqn. (B.4) holds since the value of $V_{1,\mu_{\text{TS}}^k,\nu}^{\tilde{e}_k}(s_1)$ is independent of the environment $\mathcal{E}$, Eqn. (B.5) is due to the rule of the Thompson sampling policy μ_{TS}^k, Eqn. (B.6) follows from the fact that the notations $\mathbb{E}_k(\cdot)$ and $\mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}(\cdot)$ are equivalent. Below, we introduce a lemma that characterizes the difference of performance of the same policy on two different environments.

Lemma 4. *Consider any two environments e and e' with different transition kernels P_h^e and $P_h^{e'}$ but the same reward functions $\{r_h\}_{h=1}^H$. For any fixed policy (μ,ν), we have*

$$V^{e}_{1,\mu,\nu}(s_1) - V^{e'}_{1,\mu,\nu}(s_1) \tag{B.7}$$

$$= \sum_{h=1}^{H} \mathbb{E}^{e'}_{\mu,\nu}\left[\mathbb{E}_{s'\sim P^{e}_{h}(\cdot|s_h,a_h,b_h)}[V^{e}_{h+1,\mu,\nu}(s')] - \mathbb{E}_{s'\sim P^{e'}_{h}(\cdot|s_h,a_h,b_h)}[V^{e}_{h+1,\mu,\nu}(s')]\right] \tag{B.8}$$

$$= \sum_{h=1}^{H} \mathbb{E}^{e}_{\mu,\nu}\left[\mathbb{E}_{s'\sim P^{e}_{h}(\cdot|s_h,a_h,b_h)}[V^{e'}_{h+1,\mu,\nu}(s')] - \mathbb{E}_{s'\sim P^{e'}_{h}(\cdot|s_h,a_h,b_h)}[V^{e'}_{h+1,\mu,\nu}(s')]\right]. \tag{B.9}$$

Proof of Lemma 4. The proof is adapted from [12, Lemma D.3] with appropriate modifications for two-player zero-sum MGs. □

For notational convenience, we define

$$\Delta(\mathcal{E},\tilde{e}_k,s,a,b) \triangleq \mathbb{E}_{s'\sim P^{\mathcal{E}}_{h}(\cdot|s_h,a_h,b_h)}[V^{\mathcal{E}}_{h+1,\mu^*(\mathcal{E}),\nu}(s')] - \mathbb{E}_{s'\sim P^{\tilde{e}_k}_{h}(\cdot|s_h,a_h,b_h)}[V^{\mathcal{E}}_{h+1,\mu^*(\mathcal{E}),\nu}(s')] \tag{B.10}$$

and define the *occupancy measure* with respect to any policy (μ,ν) and environment e as

$$d^{e}_{h,\mu,\nu}(s,a,b) \triangleq \mathbb{P}^{e}_{\mu,\nu}(s^k_h = s, a^k_h = a, b^k_h = b). \tag{B.11}$$

Applying Lemma 4 to Eqn. (B.6), we have

$$\left|\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\tilde{e}_k}_{1,\mu^*(\mathcal{E}),\nu}(s_1)\right]\right| \tag{B.12}$$

$$= \left|\mathbb{E}_k\left[\sum_{h=1}^{H}\mathbb{E}^{\tilde{e}_k}_{\mu^*(\mathcal{E}),\nu}\left[\Delta(\mathcal{E},\tilde{e}_k,s^k_h,a^k_h,b^k_h)\right]\right]\right| \tag{B.13}$$

$$\le \sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)\cdot\left|\Delta(\mathcal{E},\tilde{e}_k,s,a,b)\right|\right]. \tag{B.14}$$

A key observation is that the expected occupancy measure $\mathbb{E}_k[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)]$ with respect to the random environment $\mathcal{E}$ is equal to the occupancy measure of the Thompson sampling policy $d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b)$, due to the definition of μ^k_{TS}. Thus, one can further express Eqn. (B.14) as

$$\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)\cdot\left|\Delta(\mathcal{E},\tilde{e}_k,s,a,b)\right|\right] \tag{B.15}$$

$$= \sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b}\frac{d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)}{\sqrt{\mathbb{E}_k[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)]}}\cdot\sqrt{d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b)}\cdot\left|\Delta(\mathcal{E},\tilde{e}_k,s,a,b)\right|\right] \tag{B.16}$$

$$\le \sqrt{\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b}\frac{\left[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)\right]^2}{\mathbb{E}_k[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)]}\right]}\cdot\sqrt{\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b)\Delta(\mathcal{E},\tilde{e}_k,s,a,b)^2\right]} \tag{B.17}$$

$$\le \sqrt{SABH}\cdot\sqrt{\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b)\Delta(\mathcal{E},\tilde{e}_k,s,a,b)^2\right]}, \tag{B.18}$$

where (B.17) follows from the Cauchy-Schwarz inequality, and (B.18) is due to the linearity of expectation as well as the fact that the occupancy measure always satisfies $[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)]^2 \le d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)$. For the second part of (B.18), we have

$$\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b)\Delta(\mathcal{E},\tilde{e}_k,s,a,b)^2\right] \tag{B.19}$$

$$= \sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b)\left[\sum_{s'}\left(P^{\mathcal{E}}_h(s'|s,a,b) - P^{\tilde{e}_k}_h(s'|s,a,b)\right)\cdot V^{\mathcal{E}}_{h+1,\mu^*(\mathcal{E}),\nu}(s')\right]^2\right] \tag{B.20}$$

$$\le H^2\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b)\left[\sum_{s'}\left|P^{\mathcal{E}}_h(s'|s,a,b) - P^{\tilde{e}_k}_h(s'|s,a,b)\right|\right]^2\right] \tag{B.21}$$

$$= H^2 \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b) \left[\mathbb{D}_{\mathrm{TV}} \left(P^{\mathcal{E}}_h(\cdot|s,a,b), P^{\tilde{e}_k}_h(\cdot|s,a,b) \right) \right]^2 \right] \tag{B.22}$$

$$\leq H^2 \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu^k_{\mathrm{TS}},\nu}(s,a,b) \frac{1}{2} \mathbb{D}_{\mathrm{KL}} \left(P^{\mathcal{E}}_h(\cdot|s,a,b) \| P^{\tilde{e}_k}_h(\cdot|s,a,b) \right) \right], \tag{B.23}$$

where Eqn. (B.21) holds since the value function is always bounded from above by H, Eqn. (B.22) follows from the definition of the *total variation distance* $\mathbb{D}_{\mathrm{TV}}(\cdot,\cdot)$ between two distributions, and in Eqn. (B.23) we apply the Pinsker's inequality to relate the total variation distance to the KL-divergence between two distributions. Finally, we introduce a lemma (adapted from [12, Lemma A.1]) showing that the term in (B.23) can be related to the mutual information between the environment $\mathcal{E}$ and the trajectory of episode k.

Lemma 5. *For any policy* (μ,ν), *we have*

$$\mathbb{I}^{\mu,\nu}_k(\mathcal{E}; \mathcal{T}^k_{H+1}) = \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a,b} d^{\tilde{e}_k}_{h,\mu,\nu}(s,a,b) \cdot \mathbb{D}_{\mathrm{KL}} \left(P^{\mathcal{E}}_h(\cdot|s,a,b) \| P^{\tilde{e}_k}_h(\cdot|s,a,b) \right) \right]. \tag{B.24}$$

Proof of Lemma 5. Recall that $\mathcal{T}^k_{H+1}$ is the trajectory at episode k. We further define

$$\mathcal{T}^k_h \triangleq \{s^k_1, a^k_1, b^k_1, r^k_1, \dots, s^k_h, a^k_h, b^k_h, r^k_h\}$$

as the *trajectory* at episode k from steps 1 to h (for any $h \in [H]$). By the chain rule of mutual information, we have

$$\mathbb{I}^{\mu,\nu}_k(\mathcal{E}; \mathcal{T}^k_{H+1}) \tag{B.25}$$

$$= \sum_{h=1}^{H} \mathbb{I}^{\mu,\nu}_k(\mathcal{E}; (s^k_h, a^k_h, b^k_h, r^k_h)|\mathcal{T}^k_{h-1}) + \mathbb{I}^{\mu,\nu}_k(\mathcal{E}; s^k_{H+1}|\mathcal{T}^k_H) \tag{B.26}$$

$$= \sum_{h=1}^{H+1} \mathbb{I}^{\mu,\nu}_k(\mathcal{E}; s^k_h|\mathcal{T}^k_{h-1}) + \sum_{h=1}^{H} \left[\mathbb{I}^{\mu,\nu}_k(\mathcal{E}; (a^k_h, b^k_h)|\mathcal{T}^k_{h-1}, s^k_h) + \mathbb{I}^{\mu,\nu}_k(\mathcal{E}; r^k_h|\mathcal{T}^k_{h-1}, s^k_h, a^k_h, b^k_h) \right]. \tag{B.27}$$

Note that $\mathbb{I}^{\mu,\nu}_k(\mathcal{E}; r^k_h|\mathcal{T}^k_{h-1}, s^k_h, a^k_h, b^k_h) = 0$ for all $h \in [H]$ since r^k_h is a deterministic function of (s^k_h, a^k_h, b^k_h). Moreover, one can also show that $\mathbb{I}^{\mu,\nu}_k(\mathcal{E}; (a^k_h, b^k_h)|\mathcal{T}^k_{h-1}, s^k_h) = 0$ for all $h \in [H]$, since

$$\mathbb{I}^{\mu,\nu}_k(\mathcal{E}; (a^k_h, b^k_h)|\mathcal{T}^k_{h-1}, s^k_h) = \mathbb{E}_k \left[\mathbb{D}_{\mathrm{KL}} \left(P(a^k_h, b^k_h|\mathcal{E}, \mathcal{T}^k_{h-1}, s^k_h) \| P(a^k_h, b^k_h|\mathcal{T}^k_{h-1}, s^k_h) \right) \right], \tag{B.28}$$

and the fact that the actions (a^k_h, b^k_h) only depend on the policy (μ,ν) and the current state s^k_h (which implies the KL-divergence term equals zero). Then, it suffices to focus on the first term of (B.27). For each step $2 \leq h \leq H+1$, we have

$$\mathbb{I}^{\mu,\nu}_k(\mathcal{E}; s^k_h|\mathcal{T}^k_{h-1}) \tag{B.29}$$

$$= \mathbb{E}_{\mathcal{T}^k_{h-1}} \left[\mathbb{E}_{\mathcal{E}\sim P(\cdot|\mathcal{D}_k, \mathcal{T}^k_{h-1})} \left[\mathbb{D}_{\mathrm{KL}} \left(P(s^k_h|\mathcal{E}, \mathcal{T}^k_{h-1}, \mathcal{D}_k) \| P(s^k_h|\mathcal{T}^k_{h-1}, \mathcal{D}_k) \right) \right] \right] \tag{B.30}$$

$$= \mathbb{E}_{\mathcal{T}^k_{h-1}} \left[\mathbb{E}_{\mathcal{E}\sim \mathbb{P}(\cdot|\mathcal{D}_k)} \left[\mathbb{D}_{\mathrm{KL}} \left(P(s^k_h|\mathcal{E}, \mathcal{T}^k_{h-1}, \mathcal{D}_k) \| P(s^k_h|\mathcal{T}^k_{h-1}, \mathcal{D}_k) \right) \right] \right] \tag{B.31}$$

$$= \mathbb{E}_{\mathcal{T}^k_{h-1}} \left[\mathbb{E}_{\mathcal{E}\sim \mathbb{P}(\cdot|\mathcal{D}_k)} \left[\mathbb{D}_{\mathrm{KL}} \left(P^{\mathcal{E}}_{h-1}(\cdot|s^k_{h-1}, a^k_{h-1}, b^k_{h-1}) \| P^{\tilde{e}_k}_{h-1}(\cdot|s^k_{h-1}, a^k_{h-1}, b^k_{h-1}) \right) \right] \right] \tag{B.32}$$

$$= \sum_{s,a,b} \mathbb{P}^{\tilde{e}_k}_{\mu,\nu}(s^k_{h-1} = s, a^k_{h-1} = a, b^k_{h-1} = b) \cdot \mathbb{E}_k \left[\mathbb{D}_{\mathrm{KL}} \left(P^{\mathcal{E}}_{h-1}(\cdot|s,a,b) \| P^{\tilde{e}_k}_{h-1}(\cdot|s,a,b) \right) \right] \tag{B.33}$$

$$= \mathbb{E}_k \left[\sum_{s,a,b} d^{\tilde{e}_k}_{h-1,\mu,\nu}(s,a,b) \cdot \mathbb{D}_{\mathrm{KL}} \left(P^{\mathcal{E}}_{h-1}(\cdot|s,a,b) \| P^{\tilde{e}_k}_{h-1}(\cdot|s,a,b) \right) \right], \tag{B.34}$$

where (B.31) is due to the fact that the prior distribution on $\mathcal{E}$ is a product distribution over the H steps, so that the trajectory $\mathcal{T}^k_{h-1}$ does not affect the posterior distribution of $\mathcal{E}$ with respect to the transition kernel P_{h-1} appeared in the KL-divergence term. Eqn. (B.32) holds since (i) the distribution of s_h only depends on $(s^k_{h-1}, a^k_{h-1}, b^k_{h-1})$ and the environment; and (ii) the definition of the mean environment $\tilde{e}_k$ ensures that

$$\begin{aligned} P(s^k_h|\mathcal{T}^k_{h-1}, \mathcal{D}_k) &= P(s^k_h|s^k_{h-1}, a^k_{h-1}, b^k_{h-1}, \mathcal{D}_k) \\ &= \mathbb{E}_{\mathcal{E}\sim \mathbb{P}(\cdot|\mathcal{D}_k)} \left[P(s^k_h|s^k_{h-1}, a^k_{h-1}, b^k_{h-1}, \mathcal{D}_k, \mathcal{E}) \right] \\ &= P^{\tilde{e}_k}_{h-1}(\cdot|s^k_{h-1}, a^k_{h-1}, b^k_{h-1}). \end{aligned} \tag{B.35}$$

Eqn. (B.33) is due to the fact that

$$\mathbb{E}_k \left(\mathbb{P}^{\mathcal{E}}_{\mu,\nu}(s^k_{h-1} = s, a^k_{h-1} = a, b^k_{h-1} = b) \right) = \mathbb{P}^{\tilde{e}_k}_{\mu,\nu}(s^k_{h-1} = s, a^k_{h-1} = a, b^k_{h-1} = b). \tag{B.36}$$

Also note that $\mathbb{I}_k^{\mu,\nu}(\mathcal{E}; s_h^k | \mathcal{T}_{h-1}^k) = 0$ when $h = 1$, as the state $s_1^k = s_1$ is deterministic. Therefore, we have

$$\mathbb{I}_k^{\mu,\nu}(\mathcal{E}; \mathcal{T}_{H+1}^k) = \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a,b} d_{h,\mu,\nu}^{\tilde{e}_k}(s,a,b) \cdot \mathbb{D}_{\mathrm{KL}} \left(P_h^{\mathcal{E}}(\cdot|s,a,b) \| P_h^{\tilde{e}_k}(\cdot|s,a,b) \right) \right]. \tag{B.37}$$

This completes the proof of Lemma 5. □

Applying Lemma 5, we have

$$\left| \mathbb{E}_k \left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1) - V_{1,\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k}(s_1) \right] \right| \le \sqrt{SABH} \cdot \sqrt{H^2 \mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_H^k)} \tag{B.38}$$

$$= \sqrt{SABH^3 \mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_H^k)}. \tag{B.39}$$

For the second term in the RHS of (B.2), we have

$$\left| \mathbb{E}_k \left[V_{1,\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k}(s_1) - V_{1,\mu_{\mathrm{TS}}^k,\nu}^{\mathcal{E}}(s_1) \right] \right| \tag{B.40}$$

$$= \left| \mathbb{E}_k \left[\sum_{h=1}^{H} \mathbb{E}_{\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k} \left[\sum_{s'} P_h^{\tilde{e}_k}(s'|s_h^k,a_h^k,b_h^k) V_{h+1,\mu_{\mathrm{TS}}^k,\nu}^{\mathcal{E}}(s') - \sum_{s'} P_h^{\mathcal{E}}(s'|s_h^k,a_h^k,b_h^k) V_{h+1,\mu_{\mathrm{TS}}^k,\nu}^{\mathcal{E}}(s') \right] \right] \right| \tag{B.41}$$

$$\le H \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a,b} d_{h,\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k}(s,a,b) \left[\sum_{s'} \left| P_h^{\tilde{e}_k}(s'|s,a,b) - P_h^{\mathcal{E}}(s'|s,a,b) \right| \right] \right] \tag{B.42}$$

$$\le H \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a,b} d_{h,\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k}(s,a,b) \sqrt{\frac{1}{2} \mathbb{D}_{\mathrm{KL}} \left(P_h^{\mathcal{E}}(\cdot|s,a,b) \| P_h^{\tilde{e}_k}(\cdot|s,a,b) \right)} \right] \tag{B.43}$$

$$\le H^2 \sqrt{\frac{1}{2H} \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a,b} d_{h,\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k}(s,a,b) \cdot \mathbb{D}_{\mathrm{KL}} \left(P_h^{\mathcal{E}}(\cdot|s,a,b) \| P_h^{\tilde{e}_k}(\cdot|s,a,b) \right) \right]} \tag{B.44}$$

$$= \sqrt{\frac{H^3}{2} \mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_{H+1}^k)}, \tag{B.45}$$

where (B.41) is due to Lemma 4, (B.42) is due to the fact that the value function is always bounded from above by H, (B.43) follows from the Pinsker's inequality, (B.44) follows from Jensen's inequality, and (B.45) follows from Lemma 5.

Therefore, we obtain

$$\Gamma_k(\mu_{\mathrm{TS}}^k, \nu, \mathcal{E}) = \frac{\left(\mathbb{E}_k \left[V_{1,\mu^*(\mathcal{E}),\nu}^{\mathcal{E}}(s_1) - V_{1,\mu_{\mathrm{TS}}^k,\nu}^{\mathcal{E}}(s_1) \right] \right)^2}{\mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_{H+1}^k)} \tag{B.46}$$

$$\le \frac{\left(\left| \mathbb{E}_k \left[V_1^{\mathcal{E},*}(s_1) - V_{1,\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k}(s_1) \right] \right| + \left| \mathbb{E}_k \left[V_{1,\mu_{\mathrm{TS}}^k,\nu}^{\tilde{e}_k}(s_1) - V_{1,\mu_{\mathrm{TS}}^k,\nu}^{\mathcal{E}}(s_1) \right] \right| \right)^2}{\mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_{H+1}^k)} \tag{B.47}$$

$$\le \frac{\left(\sqrt{SABH^3 \mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_{H+1}^k)} + \sqrt{\frac{H^3}{2} \mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_{H+1}^k)} \right)^2}{\mathbb{I}_k^{\mu_{\mathrm{TS}}^k,\nu}(\mathcal{E}; \mathcal{T}_{H+1}^k)} \tag{B.48}$$

$$= 4SABH^3. \tag{B.49}$$

This completes the proof of Lemma 1.

B.2. *Proof for the second term in* (19)

Following similar steps in Eqns. (20)-(24) and Eqns. (27)-(30), we have

$$\mathbb{E}\left(\sum_{k=1}^{K} V_{1,\mu_{\mathrm{IDS}}^k,\nu_{\mathrm{IDS}}^k}^{\mathcal{E}}(s_1) - V_{1,\mu_{\mathrm{IDS}}^k,\dagger}^{\mathcal{E}}(s_1) \right) \tag{B.50}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}\left(V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\dagger}(s_1)\right)\right] \tag{B.51}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\sqrt{\frac{\left(\mathbb{E}_k\left(V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\dagger}(s_1)\right)\right)^2}{\mathbb{I}_k^{\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(\mathcal{E};\mathcal{T}^k_{H+1})}}\sqrt{\mathbb{I}_k^{\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(\mathcal{E};\mathcal{T}^k_{H+1})}\right] \tag{B.52}$$

$$\le \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\frac{\left(\mathbb{E}_k\left(V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\dagger}(s_1)\right)\right)^2}{\mathbb{I}_k^{\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(\mathcal{E};\mathcal{T}^k_{H+1})}\right]} \cdot \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(\mathcal{E};\mathcal{T}^k_{H+1})\right]} \tag{B.53}$$

$$\le \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\Lambda_k^{\mu^k_{\mathrm{IDS}}}(\nu^k_{\mathrm{IDS}},\mathcal{E})\right]} \cdot \sqrt{\mathbb{I}^{\mu_{\mathrm{IDS}},\nu_{\mathrm{IDS}}}(\mathcal{E};\mathcal{D}^{K+1})}. \tag{B.54}$$

Recall that for each episode k and for every possible trajectory $\mathcal{D}_k$, the max-player's policy μ^k_{IDS} can be calculated based on the minimax optimization in (13). Let $\nu^k_{\mathrm{TS}}(\mu^k_{\mathrm{IDS}})$ be the TS policy of the min-player with respect to the max-player's policy μ^k_{IDS}, such that it first samples a realization of the environment $\mathcal{E}=e$ according to $\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)$, and then chooses the best response to μ^k_{IDS} under environment e (which is denoted by $\nu^\dagger_e(\mu^k_{\mathrm{IDS}})$). When there is no confusion, we may use the abbreviation ν^k_{TS} to replace $\nu^k_{\mathrm{TS}}(\mu^k_{\mathrm{IDS}})$ for simplicity. Note that the TS policy ν^k_{TS} is a Markov policy. By the definition of ν^k_{IDS}, we have

$$\Lambda_k^{\mu^k_{\mathrm{IDS}}}(\nu^k_{\mathrm{IDS}},\mathcal{E}) \le \Lambda_k^{\mu^k_{\mathrm{IDS}}}(\nu^k_{\mathrm{TS}}(\mu^k_{\mathrm{IDS}}),\mathcal{E}). \tag{B.55}$$

Moreover, one can prove an analogous version of Lemma 1 to show that the TS policy ν^k_{TS} satisfies

$$\Lambda_k^{\mu^k_{\mathrm{IDS}}}(\nu^k_{\mathrm{TS}}(\mu^k_{\mathrm{IDS}}),\mathcal{E}) \le 4H^3SAB.$$

Combining the upper bound on $\Lambda_k^{\mu^k_{\mathrm{IDS}}}(\nu^k_{\mathrm{IDS}},\mathcal{E})$ with Lemma 2, one can eventually show that

$$\mathbb{E}\left(\sum_{k=1}^{K} V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\mathrm{IDS}},\dagger}(s_1)\right) \le 4S^{3/2}ABH^2\sqrt{K\log(SKH)}. \tag{B.56}$$

Appendix C. Proof of Theorem 2

For any policy $(\mu,\nu) = (\{\mu^k\}_{k\in[K]},\{\nu^k\}_{k\in[K]})$, we have

$$\mathsf{BR}_K(\mu) = \mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k,\dagger}(s_1)\right) \tag{C.1}$$

$$= \mathbb{E}\left(\sum_{k=K_0+1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k,\dagger}(s_1)\right) + HK_0 \tag{C.2}$$

$$= \mathbb{E}\left(\sum_{k=K_0+1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k,\nu^k}(s_1)\right) + \mathbb{E}\left(\sum_{k=K_0+1}^{K} V^{\mathcal{E}}_{1,\mu^k,\nu^k}(s_1) - V^{\mathcal{E}}_{1,\mu^k,\dagger}(s_1)\right) + HK_0. \tag{C.3}$$

Similar to Eqns. (20)-(24), we can upper-bound the first term in (C.3) as

$$\sqrt{\sum_{k=K_0+1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\Gamma_k(\mu^k,\nu^k,\mathcal{E})\right]} \cdot \sqrt{\sum_{k=K_0+1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k,\nu^k}(\mathcal{E};\mathcal{T}^k_{H+1})\right]} \tag{C.4}$$

Now we focus on the policy pair $(\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}})$. Similar to (25) and (27)-(30) and by Lemma 2, it holds that

$$\sum_{k=K_0+1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(\mathcal{E};\mathcal{T}^k_{H+1})\right] \le \mathbb{I}^{\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}}}(\mathcal{E};\mathcal{D}_{K+1}) \le 2S^2ABH\log(SKH), \tag{C.5}$$

$$\sum_{k=K_0+1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\Gamma_k(\mu^k_{\mathrm{IDS}},\nu^k_{\mathrm{IDS}},\mathcal{E})\right] \le \sum_{k=K_0+1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu\in\Pi_B}\Gamma_k(\mu^k_{\mathrm{TS}},\nu,\mathcal{E})\right] \le (K-K_0)\max_{K_0+1\le k\le K}\mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu\in\Pi_B}\Gamma_k(\mu^k_{\mathrm{TS}},\nu,\mathcal{E})\right]. \tag{C.6}$$

Lemma 6. *When Assumption 1 holds,* $\lim_{k\to\infty} \mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu\in\Pi_B} \Gamma_k(\mu^k_{\mathrm{TS}}, \nu, \mathcal{E})\right] \le 4H^3$.

Proof of Lemma 6. The proof is largely inspired by [46]. Moreover, we remark that many steps in the proof are the same with Appendix B.1, thus we begin at the point where the two approaches diverge. Recall from Appendix B.1 that, to obtain Eqn. (B.18), we use a simple inequality

$$\sqrt{\sum_{h=1}^{H} \mathbb{E}_k\left[\sum_{s,a,b} \frac{\left[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)\right]^2}{\mathbb{E}_k[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)]}\right]} \le \sqrt{SABH}. \tag{C.7}$$

In the following, we use a refined technique to derive a tighter upper bound for (C.7). First, we define

$$\mathcal{G}^\nu_k \triangleq \left\{(s,a,b,h) : \mathbb{E}_k[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)] \ne 0\right\}, \quad \text{and } T^\nu_k \triangleq \sum_{(s,a,b,h)\in\mathcal{G}^\nu_k} \frac{\mathbb{E}_k\left[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)^2\right]}{\mathbb{E}_k[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)]}.$$

Let $\mathcal{E}'$ be a random variable with the same distribution of $\mathcal{E}$ but is independent of $\mathcal{E}$. One can verify that

$$\mathbb{E}_k[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)] = \mathbb{E}_k[d^{\mathcal{E}'}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)], \quad \text{and } \mathbb{E}_k\left[d^{\tilde{e}_k}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)^2\right] \le \mathbb{E}_k\left[d^{\mathcal{E}'}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)^2\right]. \tag{C.8}$$

Let $g^\nu_k(s,a,b,h,\mathcal{D}_k) \triangleq \mathbb{E}_k\left[d^{\mathcal{E}'}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)^2\right] / \mathbb{E}_k[d^{\mathcal{E}'}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)]$ if $(s,a,b,h)\in\mathcal{G}^\nu_k$, and $g^\nu_k(s,a,b,h,\mathcal{D}_k) \triangleq 0$ otherwise. Then, we have

$$T^\nu_k \le \sum_{(s,a,b,h)\in\mathcal{G}^\nu_k} g^\nu_k(s,a,b,h,\mathcal{D}_k).$$

Suppose the ground truth environment is $\mathcal{E}_0$. According to Assumption 1 and [46, Corollary 12], we have that for almost every $\mathcal{D}_k$ sampled from $\mathcal{E}_0$,

$$\lim_{k\to\infty} \mathbb{E}_k[d^{\mathcal{E}'}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)] = d^{\mathcal{E}_0}_{h,\mu^*(\mathcal{E}_0),\nu}(s,a,b), \tag{C.9}$$

$$\lim_{k\to\infty} \mathbb{E}_k[d^{\mathcal{E}'}_{h,\mu^*(\mathcal{E}),\nu}(s,a,b)^2] = d^{\mathcal{E}_0}_{h,\mu^*(\mathcal{E}_0),\nu}(s,a,b)^2, \tag{C.10}$$

$$\lim_{k\to\infty} \mathbb{1}\{(s,a,b,h)\in\mathcal{G}^\nu_k\} = \mathbb{1}\{d^{\mathcal{E}_0}_{h,\mu^*(\mathcal{E}_0),\nu}(s,a,b) \ne 0\}, \tag{C.11}$$

and thus

$$\lim_{k\to\infty} T^\nu_k \le \sum_{(s,a,b,h)} d^{\mathcal{E}_0}_{h,\mu^*(\mathcal{E}_0),\nu}(s,a,b) \cdot \mathbb{1}\{d^{\mathcal{E}_0}_{h,\mu^*(\mathcal{E}_0),\nu}(s,a,b) \ne 0\} \le H. \tag{C.12}$$

Combining (B.2), (B.17), (B.23), (B.45), (C.12) and Lemma 5, we obtain that $\lim_{k\to\infty} \Gamma_k(\mu^k_{\mathrm{TS}}, \nu, \mathcal{E}) \le 4H^3$. Using the dominated convergence theorem and the boundedness of $|\Pi_B|$, we have

$$\lim_{k\to\infty} \mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu\in\Pi_B} \Gamma_k(\mu^k_{\mathrm{TS}}, \nu, \mathcal{E})\right] = \lim_{k\to\infty} \mathbb{E}_{\mathcal{E}_0}\left[\mathbb{E}_{\mathcal{D}_k\sim\mathbb{P}(\cdot|\mathcal{E}_0)}\left(\max_{\nu\in\Pi_B} \Gamma_k(\mu^k_{\mathrm{TS}}, \nu, \mathcal{E})\right)\right] \tag{C.13}$$

$$= \mathbb{E}_{\mathcal{E}_0}\left[\mathbb{E}_{\mathcal{D}_k\sim\mathbb{P}(\cdot|\mathcal{E}_0)}\left(\max_{\nu\in\Pi_B} \lim_{k\to\infty} \Gamma_k(\mu^k_{\mathrm{TS}}, \nu, \mathcal{E})\right)\right] \le 4H^3. \tag{C.14}$$

This completes the proof of Lemma 6. □

Based on Lemma 6, we know that there must exist a constant K_0 such that

$$\max_{K_0+1\le k\le K} \mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu\in\Pi_B} \Gamma_k(\mu^k_{\mathrm{TS}}, \nu, \mathcal{E})\right] \le 4H^3.$$

Thus, combining (C.4)-(C.6), we can upper-bound the first term in (C.3) as

$$\mathbb{E}\left(\sum_{k=K_0+1}^{K} V^{\mathcal{E},*}_1(s_1) - V^{\mathcal{E}}_{1,\mu^k,\nu^k}(s_1)\right) \le \sqrt{4KH^3} \cdot \sqrt{2S^2ABH\log(SKH)} = \sqrt{8S^2ABH^4K\log(SKH)}. \tag{C.15}$$

The second term in (C.3) can be upper-bounded in a similar manner. This completes the proof of Theorem 2.

Appendix D. Proof of Theorem 3 (Reg-MAIDS)

We first focus on the first term in (19). Note that for any min-player's policy $\nu = \{\nu^k\}_{k=1}^K$, the max-player's policy $\mu_{\text{R-IDS}} = \{\mu^k_{\text{R-IDS}}\}_{k=1}^K$ satisfies

$$
\begin{aligned}
&\mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\nu^k}(s_1)\right)\\
&= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left(V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\nu^k}(s_1)\right)\right]\\
&= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left[V_1^{\mathcal{E},*}(s_1)\right] - \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\nu^k}(s_1)\right] - \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu^k}(\mathcal{E};\mathcal{T}^k_{H+1}) + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu^k}(\mathcal{E};\mathcal{T}^k_{H+1})\right]\\
&\le \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left[V_1^{\mathcal{E},*}(s_1)\right] - \left(\min_{\nu'} \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\nu'}(s_1)\right] + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})\right) + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu^k}(\mathcal{E};\mathcal{T}^k_{H+1})\right].
\end{aligned}
\tag{D.1}
$$

By the definition of $\mu^k_{\text{R-IDS}}$, we have

$$
\min_{\nu'} \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\nu'}(s_1)\right] + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1}) \ge \min_{\nu'} \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^k_{\text{IDS}},\nu'}(s_1)\right] + \lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1}).
$$

Thus, (D.1) can be bounded from above as

$$
\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left[V_1^{\mathcal{E},*}(s_1)\right] - \left(\min_{\nu'} \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\nu'}(s_1)\right] + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})\right) + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu^k}(\mathcal{E};\mathcal{T}^k_{H+1})\right]\\
&\le \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left[V_1^{\mathcal{E},*}(s_1)\right] - \left(\min_{\nu'} \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^k_{\text{IDS}},\nu'}(s_1)\right] + \lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})\right) + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu^k}(\mathcal{E};\mathcal{T}^k_{H+1})\right]\\
&= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\left(\max_{\nu'} \mathbb{E}_k\left[V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{IDS}},\nu'}(s_1)\right] - \lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})\right) + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu^k}(\mathcal{E};\mathcal{T}^k_{H+1})\right].
\end{aligned}
$$

Next, due to the AM-GM inequality, we have

$$
\begin{aligned}
&\mathbb{E}_k\left[V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{IDS}},\nu'}(s_1)\right]\\
&\le \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu'}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{IDS}},\nu'}(s_1)\right]\\
&= \frac{\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu'}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{IDS}},\nu'}(s_1)\right]}{\sqrt{\lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})}} \cdot \sqrt{\lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})}\\
&\le \frac{\left(\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu'}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{IDS}},\nu'}(s_1)\right]\right)^2}{4\lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})} + \lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1})\\
&= \frac{1}{4\lambda}\Gamma_k(\mu^k_{\text{IDS}},\nu',\mathcal{E}) + \lambda \mathbb{I}_k^{\mu^k_{\text{IDS}},\nu'}(\mathcal{E};\mathcal{T}^k_{H+1}).
\end{aligned}
$$

Therefore, for the policy $(\mu_{\text{R-IDS}},\nu_{\text{R-IDS}})$ of REG-MAIDS, we have

$$
\mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\nu^k_{\text{R-IDS}}}(s_1)\right) \tag{D.2}
$$

$$
\le \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\left(\max_{\nu'} \frac{\Gamma_k(\mu^k_{\text{IDS}},\nu',\mathcal{E})}{4\lambda}\right) + \lambda \mathbb{I}_k^{\mu^k_{\text{R-IDS}},\nu^k_{\text{R-IDS}}}(\mathcal{E};\mathcal{T}^k_{H+1})\right] \tag{D.3}
$$

$$
\le \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\left(\max_{\nu'} \frac{\Gamma_k(\mu^k_{\text{TS}},\nu',\mathcal{E})}{4\lambda}\right)\right] + \lambda \mathbb{I}^{\mu_{\text{R-IDS}},\nu_{\text{R-IDS}}}(\mathcal{E};\mathcal{D}_{K+1}) \tag{D.4}
$$

$$
\le K \cdot \frac{4H^3 SAB}{4\lambda} + \lambda \cdot 2S^2 ABH \log(SKH) \tag{D.5}
$$

$$
= \frac{5}{2}\sqrt{2H^4 S^3 A^2 B^2 K \log(SKH)}, \tag{D.6}
$$

where Eqn. (D.4) follows from the definition of μ^k_{IDS} in (13) as well as the derivations in (27)-(30). Eqn. (D.5) is obtained by recalling the upper bound on $\Gamma_k(\mu^k_{\text{TS}}, \nu, \mathcal{E})$ presented in Lemma 1 and the upper bound on $\mathbb{I}^{\mu_{\text{R-IDS}},\nu}(\mathcal{E}; \mathcal{D}_{K+1})$ presented in Lemma 2, while Eqn. (D.6) is obtained by choosing

$$\lambda = \sqrt{2KH^2/S\log(SKH)}.$$

We remark that the second term in (19) can be handled in a similar fashion. By choosing the parameter

$$\widetilde{\lambda} = \lambda = \sqrt{2KH^2/S\log(SKH)},$$

one can show that the second term in (19) is also upper bounded by $\frac{5}{2}\sqrt{2H^4S^3A^2B^2K\log(SKH)}$, leading to the following upper bound on the Bayesian regret:

$$\mathsf{BR}_K(\mu_{\text{R-IDS}}) = \mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{R-IDS}},\dagger}(s_1)\right) \le 8H^2S^{3/2}AB\sqrt{K\log(SKH)}. \tag{D.7}$$

Appendix E. Appendix for Compressed-MAIDS

E.1. Comparison of distortion measures

Remark 4. The distortion measure d_{Φ_A,Φ_B} is similar to the one defined in [22] for MDPs, which follows the *value equivalence principle* [48,49]. We note that their definition focuses on the value difference through the Bellman update; specifically, for some value function class $\mathcal{V}$ and two environments e and e', they define the distortion between e and e' by

$$\sup_{\mu\in\Phi_A}\sup_{V\in\mathcal{V}}\left|\mathbb{E}_{a\sim\mu(\cdot|s),s'\sim P^e(\cdot|s,a)}[V(s')] - \mathbb{E}_{a\sim\mu(\cdot|s),s'\sim P^{e'}(\cdot|s,a)}[V(s')]\right|$$

in the context of time-homogeneous MDPs with only one player. Note that they require an additional value function class $\mathcal{V}$; in contrast, we directly let the value function $V^e_{1,\mu,\nu}(s_1)$ depend on the policy and the environment. The similarity of these two measures (if considered in the context of zero-sum MGs), roughly speaking, is that our definition $|V^e_{1,\mu,\nu}(s_1) - V^{e'}_{1,\mu,\nu}(s_1)|$ can be seen as the *cumulative* differences of values through the Bellman update for the H steps.

E.2. Lemma 7 and its proof

Lemma 7 below shows that the construction of $\tilde{\mathcal{E}}$ in (41) in Section 6.2 satisfies the hard-compression constraint in (39).

Lemma 7. *For every episode $k \in [K]$ and any $\Phi_A \subseteq \Pi_A$ and $\Phi_B \subseteq \Pi_B$, the compressed environment $\tilde{\mathcal{E}}$ satisfies that*

$$\mathbb{P}(d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) > \epsilon) = 0.$$

Proof. For any subspace Θ_c, any pair of environment $e, e' \in \Theta_c$, and any policy $(\mu,\nu) \in \Phi_A \times \Phi_B$, we have

$$V^e_{1,\mu,\nu}(s_1) - V^{e'}_{1,\mu,\nu}(s_1) \tag{E.1}$$

$$= \sum_{h=1}^{H} \mathbb{E}^{e'}_{\mu,\nu}\left[\mathbb{E}_{s'\sim P^e_h(\cdot|s_h,a_h,b_h)}[V^e_{h+1,\mu,\nu}(s')] - \mathbb{E}_{s'\sim P^{e'}_h(\cdot|s_h,a_h,b_h)}[V^e_{h+1,\mu,\nu}(s')\right] \tag{E.2}$$

$$= \sum_{h=1}^{H} \mathbb{E}^{e'}_{\mu,\nu}\left[[P^e_h(\cdot|s_h,a_h,b_h) - P^{e'}_h(\cdot|s_h,a_h,b_h)]\cdot V^e_{h+1,\mu,\nu}(s')\right] \tag{E.3}$$

$$\le H\sum_{h=1}^{H} \mathbb{E}^{e'}_{\mu,\nu}\left[\left|P^e_h(\cdot|s_h,a_h,b_h) - P^{e'}_h(\cdot|s_h,a_h,b_h)\right|\right] \tag{E.4}$$

$$\le H\sum_{h=1}^{H}\max_{s,a,b}\left|P^e_h(\cdot|s,a,b) - P^{e'}_h(\cdot|s,a,b)\right| \tag{E.5}$$

$$\le \epsilon, \tag{E.6}$$

where (E.2) follows from Lemma 4, and the last inequality follows from the property of Θ_c. Therefore, based on the construction of $\tilde{\mathcal{E}}$ in (41), it is clear that

$$\mathbb{P}\left(d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) > \epsilon\right) = 0. \tag{E.7}$$

This completes the proof of Lemma 7. □

E.3. Proof of Theorem 4

Recall that

$$\mathsf{BR}_K(\mu_{\text{C-IDS}}) = \mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\dagger}(s_1)\right) \tag{E.8}$$

$$= \mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1)\right) + \mathbb{E}\left(\sum_{k=1}^{K} V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\dagger}(s_1)\right). \tag{E.9}$$

For the first term in (E.9), we have

$$\mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1)\right) \tag{E.10}$$

$$\leq \mathbb{E}\left(\sum_{k=1}^{K} V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1)\right) \tag{E.11}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left(V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1)\right) - 2\epsilon\right] + 2K\epsilon. \tag{E.12}$$

In Lemma 8 below, we characterize the difference of performance of the same policy on the environment $\mathcal{E}$ and the compressed environment $\tilde{\mathcal{E}}$.

Lemma 8. *For any episode $k \in [K]$ and any policy (μ, ν), we have*

$$\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\mathcal{E}}_{1,\mu,\nu}(s_1)\right] - 2\epsilon \leq \mathbb{E}_k\left[V^{\tilde{\mathcal{E}}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu,\nu}(s_1)\right]. \tag{E.13}$$

Proof of Lemma 8. Recall that the compressed environment $\tilde{\mathcal{E}}$ constructed in (41) satisfies the hard-compression constraint for any $\Phi_A \subseteq \Pi_A$ and $\Phi_B \subseteq \Pi_B$. Setting $\Phi_A = \Pi_A$ and $\Phi_B = \Pi_B$, we have

$$\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu}(s_1)\right] - \mathbb{E}_k\left[V^{\tilde{\mathcal{E}}}_{1,\mu^*(\mathcal{E}),\nu}(s_1)\right] \tag{E.14}$$

$$= \mathbb{P}_k(d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) > \epsilon) \cdot \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) \middle| d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) > \epsilon\right]$$

$$+ \mathbb{P}_k(d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) \leq \epsilon) \cdot \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) \middle| d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) \leq \epsilon\right] \tag{E.15}$$

$$= \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu^*(\mathcal{E}),\nu}(s_1) \middle| d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) \leq \epsilon\right] \tag{E.16}$$

$$\leq \epsilon, \tag{E.17}$$

where Eqn. (E.16) is due to the fact that the compressed environment $\tilde{\mathcal{E}}$ satisfies $\mathbb{P}_k(d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}}) > \epsilon) = 0$, and Eqn. (E.17) follows from the definition of $d_{\Phi_A,\Phi_B}(\mathcal{E},\tilde{\mathcal{E}})$ for $\Phi_A = \Pi_A$ and $\Phi_B = \Pi_B$. Similarly, one can also show that

$$\mathbb{E}_k\left[V^{\tilde{\mathcal{E}}}_{1,\mu,\nu}(s_1)\right] - \mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu,\nu}(s_1)\right] \leq \epsilon. \tag{E.18}$$

Combining (E.17) and (E.18) together, we complete the proof of Lemma 8. $\square$

Based on Lemma 8, we can bound (E.12) from above as

$$\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left(V^{\mathcal{E}}_{1,\mu^*(\mathcal{E}),\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1)\right) - 2\epsilon\right] + 2K\epsilon \tag{E.19}$$

$$\leq \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left(V^{\tilde{\mathcal{E}}}_{1,\mu^*(\mathcal{E}),\nu^k_{\text{C-IDS}}}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1)\right)\right] + 2K\epsilon \tag{E.20}$$

$$\leq \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\sqrt{\tilde{\Gamma}_k(\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}},\tilde{\mathcal{E}})}\sqrt{\mathbb{I}_k^{\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(\tilde{\mathcal{E}};\mathcal{T}^k_{H+1})}\right] + 2K\epsilon \tag{E.21}$$

$$\leq \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\tilde{\Gamma}_k(\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}},\tilde{\mathcal{E}})\right]} \cdot \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(\tilde{\mathcal{E}};\mathcal{T}^k_{H+1})\right]} + 2K\epsilon \tag{E.22}$$

$$\leq \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu} \tilde{\Gamma}_k(\mu^k_{\text{C-IDS}}, \nu, \tilde{\mathcal{E}})\right]} \cdot \sqrt{\mathbb{I}^{\mu_{\text{C-IDS}}, \nu_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{D}_{K+1})} + 2K\epsilon \tag{E.23}$$

$$\leq \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\max_{\nu} \tilde{\Gamma}_k(\mu^k_{\text{TS}}, \nu, \tilde{\mathcal{E}})\right]} \cdot \sqrt{\mathbb{I}^{\mu_{\text{C-IDS}}, \nu_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{D}_{K+1})} + 2K\epsilon, \tag{E.24}$$

where (E.22) follows from the Cauchy-Schwarz inequality, and (E.24) is due to the definition of the max-player's policy $\mu_{\text{C-IDS}}$, as well as the chain rule of mutual information. Using the same proof technique as for Lemma 1, one can prove that for any ν and any distribution of $\mathcal{E}$,

$$\tilde{\Gamma}_k(\mu^k_{\text{TS}}, \nu, \tilde{\mathcal{E}}) \leq 4H^3SAB. \tag{E.25}$$

This means that the first term in (E.9) satisfies

$$\mathbb{E}\left(\sum_{k=1}^{K} V_1^{\mathcal{E},*}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1)\right) \leq \sqrt{4KH^3SAB \cdot \mathbb{I}^{\mu_{\text{C-IDS}}, \nu_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{D}_{K+1})} + 2K\epsilon. \tag{E.26}$$

For the second term in (E.9), we have

$$\mathbb{E}\left(\sum_{k=1}^{K} V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\dagger}(s_1)\right) \tag{E.27}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left(V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\dagger}(s_1)\right) - 2\epsilon\right] + 2K\epsilon. \tag{E.28}$$

Similar to Lemma 8, we can also show that

$$\mathbb{E}_k\left[V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^\dagger_{\mathcal{E}}(\mu^k_{\text{C-IDS}})}(s_1)\right] - 2\epsilon$$
$$\leq \mathbb{E}_k\left[V^{\tilde{\mathcal{E}}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu^k_{\text{C-IDS}},\nu^\dagger_{\tilde{\mathcal{E}}}(\mu^k_{\text{C-IDS}})}(s_1)\right]. \tag{E.29}$$

Thus, (E.28) can be bounded from above as

$$\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{E}_k\left(V^{\tilde{\mathcal{E}}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1) - V^{\tilde{\mathcal{E}}}_{1,\mu^k_{\text{C-IDS}},\nu^\dagger_{\tilde{\mathcal{E}}}(\mu^k_{\text{C-IDS}})}(s_1)\right)\right] + 2K\epsilon \tag{E.30}$$

$$\leq \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\sqrt{\tilde{\Lambda}_k^{\mu^k_{\text{C-IDS}}}(\nu^k_{\text{C-IDS}}, \tilde{\mathcal{E}})} \times \sqrt{\mathbb{I}_k^{\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{T}^k_{H+1})}\right] + 2K\epsilon \tag{E.31}$$

$$\leq \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\tilde{\Lambda}_k^{\mu^k_{\text{C-IDS}}}(\nu^k_{\text{C-IDS}}, \tilde{\mathcal{E}})\right]} \times \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{I}_k^{\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{T}^k_{H+1})\right]} + 2K\epsilon \tag{E.32}$$

$$\leq \sqrt{\sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k}\left[\tilde{\Lambda}_k^{\mu^k_{\text{C-IDS}}}(\nu^k_{\text{TS}}(\mu^k_{\text{C-IDS}}), \tilde{\mathcal{E}})\right]} \times \sqrt{\mathbb{I}^{\mu_{\text{C-IDS}}, \nu_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{D}_{K+1})} + 2K\epsilon, \tag{E.33}$$

where $\nu^k_{\text{TS}}(\mu^k_{\text{C-IDS}})$ in (E.33) is the TS policy of the min-player with respect to the max-player's policy $\mu^k_{\text{C-IDS}}$; specifically, it first samples a realization of the environment $\mathcal{E} = e$, and then chooses the best response to $\mu^k_{\text{C-IDS}}$ under environment e. Using the same proof technique as for Lemma 1, we have

$$\tilde{\Lambda}_k^{\mu^k_{\text{C-IDS}}}(\nu^k_{\text{TS}}(\mu^k_{\text{C-IDS}}), \tilde{\mathcal{E}}) \leq 4H^3SAB$$

for any distribution of $\mathcal{E}$. Thus, the second term in (E.9) satisfies

$$\mathbb{E}\left(\sum_{k=1}^{K} V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\nu^k_{\text{C-IDS}}}(s_1) - V^{\mathcal{E}}_{1,\mu^k_{\text{C-IDS}},\dagger}(s_1)\right) \leq \sqrt{4KH^3SAB \cdot \mathbb{I}^{\mu_{\text{C-IDS}}, \nu_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{D}_{K+1})} + 2K\epsilon. \tag{E.34}$$

Combining the upper bounds for the first term and the second term in (E.9) together, we eventually obtain that

$$\mathsf{BR}_K(\mu_{\text{C-IDS}}) \leq 4\sqrt{KH^3SAB \cdot \mathbb{I}^{\mu_{\text{C-IDS}}, \nu_{\text{C-IDS}}}(\tilde{\mathcal{E}}; \mathcal{D}_{K+1})} + 4K\epsilon. \tag{E.35}$$

This completes the proof of Theorem 4.

Appendix F. Proof of Lemma 3

By recalling the definition of $r_h^{\bar{e}_k}$ in Section 5, we have

$$\mathbb{E}_{\mu,\nu}^{\bar{e}_k}\left[\sum_{h=1}^{H} r_h^{\bar{e}_k}(s_h,a_h,b_h)\right] \tag{F.1}$$

$$= \mathbb{E}_{\mu,\nu}^{\bar{e}_k}\left[\sum_{h=1}^{H} r_h(s_h,a_h,b_h) + \lambda\mathbb{E}_k\left[\mathbb{D}_{\mathrm{KL}}(P_h^{\mathcal{E}}(\cdot|s_h,a_h,b_h)\|P_h^{\bar{e}_k}(\cdot|s_h,a_h,b_h))\right]\right] \tag{F.2}$$

$$= \sum_{h=1}^{H}\sum_{s,a,b} d_{h,\mu,\nu}^{\bar{e}_k}(s,a,b)\Big[r_h(s,a,b) + \lambda\mathbb{E}_k\left[\mathbb{D}_{\mathrm{KL}}(P_h^{\mathcal{E}}(\cdot|s,a,b)\|P_h^{\bar{e}_k}(\cdot|s,a,b))\right]\Big] \tag{F.3}$$

$$= \sum_{h=1}^{H}\sum_{s,a,b} \mathbb{E}_k[d_{h,\mu,\nu}^{\mathcal{E}}(s,a,b)]\cdot r_h(s,a,b) + \lambda\mathbb{E}_k\left[\sum_{h=1}^{H}\sum_{s,a,b} d_{h,\mu,\nu}^{\bar{e}_k}(s,a,b)\mathbb{D}_{\mathrm{KL}}(P_h^{\mathcal{E}}(\cdot|s,a,b)\|P_h^{\bar{e}_k}(\cdot|s,a,b))\right] \tag{F.4}$$

$$= \mathbb{E}_k\left[V_{1,\mu,\nu}^{\mathcal{E}}(s_1)\right] + \lambda\mathbb{I}_k^{\mu,\nu}(\mathcal{E};\mathcal{T}_{H+1}^k) \tag{F.5}$$

where (F.4) uses the fact that $d_{h,\mu,\nu}^{\bar{e}_k}(s,a,b) = \mathbb{E}_k[d_{h,\mu,\nu}^{\mathcal{E}}(s,a,b)]$ and (F.5) follows from Lemma 5.

Appendix G. Proofs of Theorem 5 (general-sum MGs)

In the following, we focus on the NE setting where the policy is $\pi_{\text{G-IDS}}^{\text{NE}} = \{\pi_{\text{G-IDS}}^{\text{NE},k}\}_{k\in[K]}$. We remark that the CCE setting can be analyzed in the same manner by simply replacing product policies to joint policies.

We first present a lemma (analogous to Lemma 3 for zero-sum MGs) that is crucial for the subsequent proofs. The proof of is omitted since it is similar to that of Lemma 3.

Lemma 9. *For any joint policy π, we have*

$$V_\pi^{(i),\hat{e}_k}(s_1) = \mathbb{E}_k\left[V_\pi^{(i),\mathcal{E}}(s_1)\right] + \lambda\mathbb{I}_k^\pi(\mathcal{E};\mathcal{T}_{H+1}^k).$$

For notational convenience, we abbreviate $\pi_{\text{G-IDS}}^{\text{NE}} = \{\pi_{\text{G-IDS}}^{\text{NE},k}\}_{k\in[K]}$ as $\pi = \{\pi^k\}_{k\in[K]}$, and abbreviate $V_{1,\pi}^{(i),\mathcal{E}}(s)$ as $V_\pi^{(i),\mathcal{E}}(s)$ when $h=1$. Moreover, we let $\mu_{\mathcal{E}}^{(i),\pi} \triangleq (\pi_{\mathcal{E}}^{(i),\dagger}, \pi^{(-i)})$.

Proof of regret bounds Recall that the Bayesian NE regret of π takes the form

$$\mathsf{BR}_K^{\mathrm{NE}}(\pi) \tag{G.1}$$

$$= \mathbb{E}_{\mathcal{E}\sim\rho}\left(\mathsf{Reg}_K^{\mathrm{NE}}(\mathcal{E},\pi)\right) \tag{G.2}$$

$$= \mathbb{E}\left(\sum_{k=1}^{K}\sum_{i=1}^{N} V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - V_{\pi^k}^{(i),\mathcal{E}}(s_1)\right), \tag{G.3}$$

$$= \sum_{k=1}^{K}\mathbb{E}_{\mathcal{D}_k}\left[\sum_{i=1}^{N}\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - V_{\pi^k}^{(i),\mathcal{E}}(s_1)\right)\right] \tag{G.4}$$

where $\mu_{\mathcal{E}}^{(i),\pi^k} = (\pi_{\mathcal{E}}^{k,(i),\dagger}, \pi^{k,(-i)})$. For any fixed $\mathcal{D}_k$ and any player $i\in[N]$, we have

$$\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - V_{\pi^k}^{(i),\mathcal{E}}(s_1)\right) \tag{G.5}$$

$$= \mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1)\right) - \left[\mathbb{E}_k\left(V_{\pi^k}^{(i),\mathcal{E}}(s_1)\right) + \lambda\mathbb{I}_k^{\pi^k}(\mathcal{E};\mathcal{T}_{H+1}^k)\right] + \lambda\mathbb{I}_k^{\pi^k}(\mathcal{E};\mathcal{T}_{H+1}^k) \tag{G.6}$$

$$= \mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1)\right) - V_{\pi^k}^{(i),\hat{e}_k}(s_1) + \lambda\mathbb{I}_k^{\pi^k}(\mathcal{E};\mathcal{T}_{H+1}^k), \tag{G.7}$$

where the last step follows from Lemma 9. As π^k can be a random policy, we sometimes write $V_{\pi^k}^{(i),\hat{e}_k}(s_1)$ as $\mathbb{E}_{\nu\sim\pi^k}[V_\nu^{(i),\hat{e}_k}(s_1)]$ for clarity. For player i, we introduce the Thompson sampling (TS) policy $\pi_{\text{TS}}^{(i)}$. The TS policy $\pi_{\text{TS}}^{(i)}$ first samples a realization of the environment $\mathcal{E}=e$ according to the distribution $\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)$, and then chooses the best response $\pi_e^{k,(i),\dagger}$ with respect to $\pi^{k,(-i)}$ in the environment e. Note that each best response $\pi_e^{k,(i),\dagger}$ is a pure policy, while the TS policy is a random policy. Since π^k is a Nash equilibrium of the MG $\hat{e}_k$, we have

$$V_{\pi^k}^{(i),\hat{e}_k}(s_1) = \mathbb{E}_{\nu\sim\pi^k}[V_\nu^{(i),\hat{e}_k}(s_1)] \geq \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}[V_\nu^{(i),\hat{e}_k}(s_1)]. \tag{G.8}$$

According to Lemma 9, we have

$$\mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}[V_\nu^{(i),\hat{e}_k}(s_1)] = \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[\mathbb{E}_k\left[V_\nu^{(i),\mathcal{E}}(s_1)\right] + \lambda \mathbb{I}_k^\nu(\mathcal{E};\mathcal{T}_{H+1}^k)\right]. \tag{G.9}$$

Substituting (G.8)-(G.9) to (G.7), we have

$$\begin{aligned}
&\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - V_{\pi^k}^{(i),\mathcal{E}}(s_1)\right) && \text{(G.10)}\\
&\leq \mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1)\right) - \left(\mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[\mathbb{E}_k\left[V_\nu^{(i),\mathcal{E}}(s_1)\right] + \lambda\mathbb{I}_k^\nu(\mathcal{E};\mathcal{T}_{H+1}^k)\right]\right) + \lambda\mathbb{I}_k^{\pi^k}(\mathcal{E};\mathcal{T}_{H+1}^k) && \text{(G.11)}\\
&= \mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\mathcal{E}}(s_1)\right]\right) - \lambda\mathbb{I}_k^{\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)}}(\mathcal{E};\mathcal{T}_{H+1}^k) + \lambda\mathbb{I}_k^{\pi^k}(\mathcal{E};\mathcal{T}_{H+1}^k) && \text{(G.12)}\\
&= \frac{\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\mathcal{E}}(s_1)\right]\right)}{\sqrt{\lambda\mathbb{I}_k^{\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)}}(\mathcal{E};\mathcal{T}_{H+1}^k)}}\cdot\sqrt{\lambda\mathbb{I}_k^{\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)}}(\mathcal{E};\mathcal{T}_{H+1}^k)}\\
&\qquad - \lambda\mathbb{I}_k^{\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)}}(\mathcal{E};\mathcal{T}_{H+1}^k) + \lambda\mathbb{I}_k^{\pi^k}(\mathcal{E};\mathcal{T}_{H+1}^k) && \text{(G.13)}\\
&\leq \frac{\left[\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\mathcal{E}}(s_1)\right]\right)\right]^2}{4\lambda\mathbb{I}_k^{\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)}}(\mathcal{E};\mathcal{T}_{H+1}^k)} + \lambda\mathbb{I}_k^{\pi^k}(\mathcal{E};\mathcal{T}_{H+1}^k), && \text{(G.14)}
\end{aligned}$$

where (G.14) follows from the AM–GM inequality. Below, we provide an upper bound on the numerator of the first term in (G.14).

Let's introduce a new environment

$$\hat{e}_k' = (H, \mathcal{S}, \mathcal{A}, \{P_h^{\hat{e}_k'}\}_{h\in[H]}, \{r_h^{(i)}\}_{h\in[H],i\in[N]}), \tag{G.15}$$

where the transition kernel $P_h^{\hat{e}_k'}(\cdot|s,a) = \mathbb{E}_{\mathcal{E}\sim\mathbb{P}(\cdot|\mathcal{D}_k)}[P_h^{\mathcal{E}}(\cdot|s,a)]$, and we point out that $\hat{e}_k'$ differs from $\hat{e}_k$ defined in Section 7 only in terms of the reward functions. Considering (G.14), we have

$$\begin{aligned}
&\left|\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\mathcal{E}}(s_1)\right]\right)\right| && \text{(G.16)}\\
&\leq \left|\mathbb{E}_k\left[V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1)\right] - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\hat{e}_k'}(s_1)\right]\right|\\
&\qquad + \left|\mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\hat{e}_k'}(s_1)\right] - \mathbb{E}_k\left[\mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\mathcal{E}}(s_1)\right]\right]\right|. && \text{(G.17)}
\end{aligned}$$

We then consider the two terms in (G.17) separately.

For the first term in (G.17), by recalling $\mu_{\mathcal{E}}^{(i),\pi^k} = (\pi_{\mathcal{E}}^{k,(i),\dagger}, \pi^{k,(-i)})$ as well as the definition of the TS policy $\pi_{\mathrm{TS}}^{(i)}$, we have

$$\begin{aligned}
&\left|\mathbb{E}_k\left[V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1)\right] - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\hat{e}_k'}(s_1)\right]\right| && \text{(G.18)}\\
&= \left|\mathbb{E}_k\left[\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\mathcal{E}}(s_1)\right]\right] - \mathbb{E}_k\left[\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\hat{e}_k'}(s_1)\right]\right]\right| && \text{(G.19)}\\
&= \left|\mathbb{E}_k\left[\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\left(V_\nu^{(i),\mathcal{E}}(s_1) - V_\nu^{(i),\hat{e}_k'}(s_1)\right)\right]\right|. && \text{(G.20)}
\end{aligned}$$

Adapting Lemma 4 to the general-sum MG setting, one can obtain that

$$V_\nu^{(i),\mathcal{E}}(s_1) - V_\nu^{(i),\hat{e}'_k}(s_1) = \sum_{h=1}^{H} \mathbb{E}_\nu^{\hat{e}_k}\left[\Delta(\mathcal{E}, \hat{e}'_k, \nu, s_h^k, a_h^k)\right], \tag{G.21}$$

$$\text{where } \Delta(\mathcal{E}, \hat{e}'_k, \nu, s, a) \triangleq \mathbb{E}_{s'\sim P_h^{\mathcal{E}}(\cdot|s,a)}[V_{h+1,\nu}^{(i),\mathcal{E}}(s')] - \mathbb{E}_{s'\sim P_h^{\hat{e}'_k}(\cdot|s,a)}[V_{h+1,\nu}^{(i),\mathcal{E}}(s')]. \tag{G.22}$$

We also define the occupancy measure with respect to any joint policy π and environment e as

$$d_{h,\pi}^e(s,a) \triangleq \mathbb{P}_\pi^e(s_h^k = s, a_h^k = a). \tag{G.23}$$

Therefore, (G.20) can be upper-bounded as

$$\left|\mathbb{E}_k\left[\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\left(V_\nu^{(i),\mathcal{E}}(s_1) - V_\nu^{(i),\hat{e}'_k}(s_1)\right)\right]\right| \tag{G.24}$$

$$= \left|\mathbb{E}_k\left[\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\left(\sum_{h=1}^{H}\mathbb{E}_\nu^{\hat{e}'_k}\left[\Delta(\mathcal{E}, \hat{e}'_k, \nu, s_h^k, a_h^k)\right]\right)\right]\right| \tag{G.25}$$

$$\leq \sum_{h=1}^{H}\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\left[\sum_{s,a} d_{h,\nu}^{\hat{e}'_k}(s,a)\cdot|\Delta(\mathcal{E}, \hat{e}'_k, \nu, s, a)|\right] \tag{G.26}$$

$$= \sum_{h=1}^{H}\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\sum_{s,a}\frac{d_{h,\nu}^{\hat{e}'_k}(s,a)}{\sqrt{\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}[d_{h,\nu}^{\hat{e}'_k}(s,a)]}}$$
$$\times\sqrt{\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}[d_{h,\nu}^{\hat{e}'_k}(s,a)]}\cdot|\Delta(\mathcal{E}, \hat{e}'_k, \nu, s, a)| \tag{G.27}$$

$$\leq \sqrt{\sum_{h=1}^{H}\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\sum_{s,a}\frac{d_{h,\nu}^{\hat{e}'_k}(s,a)^2}{\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}[d_{h,\nu}^{\hat{e}'_k}(s,a)]}}$$
$$\times\sqrt{\sum_{h=1}^{H}\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}\sum_{s,a}(\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}[d_{h,\nu}^{\hat{e}'_k}(s,a)])\Delta(\mathcal{E}, \hat{e}'_k, \nu, s, a)^2}. \tag{G.28}$$

Note that the first term in (G.28) is upper-bounded by $\sqrt{SAH}$ by using the fact that $d_{h,\nu}^{\hat{e}'_k}(s,a)^2 \leq d_{h,\nu}^{\hat{e}'_k}(s,a)$. For the second term in (G.28), we note that

$$\mathbb{E}_k\mathbb{E}_{\nu\sim(\pi_{\mathcal{E}}^{k,(i),\dagger}\times\pi^{k,(-i)})}[d_{h,\nu}^{\hat{e}'_k}(s,a)] = d_{h,(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}^{\hat{e}'_k}(s,a)$$

by the definition of the TS policy $\pi_{\mathrm{TS}}^{(i)}$. Moreover, by following the steps in (B.19)-(B.23) for zero-sum MGs, one can upper-bound the second term in (G.28) by

$$\sqrt{\frac{H^2}{2}\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a} d_{h,(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}^{\hat{e}'_k}(s,a)\cdot\mathbb{D}_{\mathrm{KL}}\left(P_h^{\mathcal{E}}(\cdot|s,a)\|P_h^{\hat{e}'_k}(\cdot|s,a)\right)\right]}. \tag{G.29}$$

Analogous to Lemma 5, one can also show that in the context of general-sum MGs,

$$\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{s,a} d_{h,(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}^{\hat{e}'_k}(s,a)\cdot\mathbb{D}_{\mathrm{KL}}\left(P_h^{\mathcal{E}}(\cdot|s,a)\|P_h^{\hat{e}'_k}(\cdot|s,a)\right)\right] = \mathbb{I}_k^{(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}(\mathcal{E};\mathcal{T}_{H+1}^k). \tag{G.30}$$

Thus, the first term in (G.17) satisfies

$$\left|\mathbb{E}_k\left[V_{\mu_{\mathcal{E}}^{(i)},\pi^k}^{(i),\mathcal{E}}(s_1)\right] - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\hat{e}'_k}(s_1)\right]\right|$$
$$\leq \sqrt{\frac{1}{2}SAH^3\cdot\mathbb{I}_k^{(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}(\mathcal{E};\mathcal{T}_{H+1}^k)}. \tag{G.31}$$

Next, we consider the second term in (G.17). Note that

$$\left|\mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\hat{e}'_k}(s_1)\right] - \mathbb{E}_k\left[\mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\mathcal{E}}(s_1)\right]\right]\right| \tag{G.32}$$

$$= \left| \mathbb{E}_k \left[\sum_{h=1}^{H} \mathbb{E}_{\nu \sim (\pi_{\mathrm{TS}}^{(i)} \times \pi^{k,(-i)})} \mathbb{E}_{\nu}^{\hat{e}'_k} \left[\sum_{s'} \left[P_h^{\hat{e}'_k}(s'|s_h^k, a_h^k) - P_h^{\mathcal{E}}(s'|s_h^k, a_h^k) \right] \cdot V_{\nu}^{(i),\mathcal{E}}(s') \right] \right] \right| \tag{G.33}$$

$$\le H \cdot \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a} d_{h,(\pi_{\mathrm{TS}}^{(i)} \times \pi^{k,(-i)})}^{\hat{e}'_k}(s,a) \cdot \sqrt{\frac{1}{2} \mathbb{D}_{\mathrm{KL}} \left(P_h^{\mathcal{E}}(\cdot|s,a) \| P_h^{\hat{e}'_k}(\cdot|s,a) \right)} \right] \tag{G.34}$$

$$\le H^2 \sqrt{\frac{1}{2H} \sum_{h=1}^{H} \mathbb{E}_k \left[\sum_{s,a} d_{h,(\pi_{\mathrm{TS}}^{(i)} \times \pi^{k,(-i)})}^{\hat{e}'_k}(s,a) \cdot \mathbb{D}_{\mathrm{KL}} \left(P_h^{\mathcal{E}}(\cdot|s,a) \| P_h^{\hat{e}'_k}(\cdot|s,a) \right) \right]} \tag{G.35}$$

$$\le \sqrt{\frac{H^3}{2} \mathbb{I}_k^{(\pi_{\mathrm{TS}}^{(i)} \times \pi^{k,(-i)})}(\mathcal{E}; \mathcal{T}_{H+1}^k)}, \tag{G.36}$$

where (G.35) follows from Jensen's inequality.

Therefore, we have

$$\left| \mathbb{E}_k \left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - \mathbb{E}_{\nu \sim (\pi_{\mathrm{TS}}^{(i)} \times \pi^{k,(-i)})} \left[V_{\nu}^{(i),\mathcal{E}}(s_1) \right] \right) \right| \le 2 \sqrt{SAH^3 \mathbb{I}_k^{(\pi_{\mathrm{TS}}^{(i)} \times \pi^{k,(-i)})}(\mathcal{E}; \mathcal{T}_{H+1}^k)}. \tag{G.37}$$

Substituting (G.37) to (G.14) yields that

$$\mathbb{E}_k \left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - V_{\pi^k}^{(i),\mathcal{E}}(s_1) \right) \le \frac{SAH^3}{\lambda} + \lambda \cdot \mathbb{I}_k^{\pi^k}(\mathcal{E}; \mathcal{T}_{H+1}^k). \tag{G.38}$$

Finally, we obtain that

$$\mathsf{BR}_K^{\mathrm{NE}}(\pi) = \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k} \left[\sum_{i=1}^{N} \mathbb{E}_k \left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - V_{\pi^k}^{(i),\mathcal{E}}(s_1) \right) \right] \tag{G.39}$$

$$= \frac{SAH^3KN}{\lambda} + \lambda N \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k} \left[\mathbb{I}_k^{\pi^k}(\mathcal{E}; \mathcal{T}_{H+1}^k) \right] \tag{G.40}$$

$$= \frac{SAH^3KN}{\lambda} + \lambda N \cdot \mathbb{I}^{\pi}(\mathcal{E}; \mathcal{D}_{K+1}) \tag{G.41}$$

$$= 3NS^{3/2}AH^2\sqrt{K \log(SKH)}, \tag{G.42}$$

where the last step is obtained by setting $\lambda = \sqrt{HK^2/S \log(SKH)}$. This completes the proof of Theorem 5.

Appendix H. Proof of Theorem 6 (general-sum MG)

The proof is essentially a combination of the techniques used for proving Theorems 4 and 5. For brevity, we only provide a proof sketch for learning NE here. First note that

$$\mathsf{BR}_K^{\mathrm{NE}}(\pi) = \mathbb{E}_{\mathcal{E} \sim \rho} \left(\mathsf{Reg}_K^{\mathrm{NE}}(\mathcal{E}, \pi) \right) \tag{H.1}$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k} \left[\sum_{i=1}^{N} \mathbb{E}_k \left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\mathcal{E}}(s_1) - V_{\pi^k}^{(i),\mathcal{E}}(s_1) \right) - 2\epsilon \right] + 2KN\epsilon \tag{H.2}$$

$$\le \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_k} \left[\sum_{i=1}^{N} \mathbb{E}_k \left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\tilde{\mathcal{E}}}(s_1) - V_{\pi^k}^{(i),\tilde{\mathcal{E}}}(s_1) \right) \right] + 2KN\epsilon, \tag{H.3}$$

where the last inequality follows from the fact that $\mathbb{P}[d(\mathcal{E}, \tilde{\mathcal{E}}) > \epsilon] = 0$ as well as the definition of $d(\cdot, \cdot)$ for the multi-player general-sum MG. Next, similar to Lemma 3, one can prove that for any π and $i \in [N]$,

$$V_{\pi}^{(i),\check{e}_k}(s_1) = \mathbb{E}_k(V_{\pi}^{(i),\tilde{\mathcal{E}}}(s_1)) + \lambda \mathbb{I}_k^{\pi}(\tilde{\mathcal{E}}; \mathcal{T}_{H+1}^k). \tag{H.4}$$

Using (H.4), we have

$$\mathbb{E}_k \left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\tilde{\mathcal{E}}}(s_1) - V_{\pi^k}^{(i),\tilde{\mathcal{E}}}(s_1) \right) = \mathbb{E}_k \left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\tilde{\mathcal{E}}}(s_1) \right) - V_{\pi^k}^{(i),\check{e}_k}(s_1) + \lambda \mathbb{I}_k^{\pi^k}(\tilde{\mathcal{E}}; \mathcal{T}_{H+1}^k), \tag{H.5}$$

Let $\pi_{\mathrm{TS}}^{(i)}$ be the TS policy as defined in Appendix G. Since π_k is the NE policy in the environment $\check{e}_k$ and π_k can be a random policy, we have

$$V_{\pi^k}^{(i),\check{e}_k}(s_1) = \mathbb{E}_{\nu\sim\pi^k}[V_\nu^{(i),\check{e}_k}(s_1)] \geq \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}[V_\nu^{(i),\check{e}_k}(s_1)] = \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[\mathbb{E}_k\left[V_\nu^{(i),\tilde{\mathcal{E}}}(s_1)\right] + \lambda \mathbb{I}_k^\nu(\tilde{\mathcal{E}};\mathcal{T}_{H+1}^k)\right]. \tag{H.6}$$

Combining (H.5)-(H.6) and following the similar procedures outlined in (G.10)-(G.14), we obtain

$$\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\tilde{\mathcal{E}}}(s_1) - V_{\pi^k}^{(i),\tilde{\mathcal{E}}}(s_1)\right) \leq \frac{\left[\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\tilde{\mathcal{E}}}(s_1) - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\tilde{\mathcal{E}}}(s_1)\right]\right)\right]^2}{4\lambda \mathbb{I}_k^{\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)}}(\tilde{\mathcal{E}};\mathcal{T}_{H+1}^k)} + \lambda \mathbb{I}_k^{\pi^k}(\tilde{\mathcal{E}};\mathcal{T}_{H+1}^k). \tag{H.7}$$

Analogous to (G.16)-(G.37) in Appendix G, one can prove that

$$\left|\mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\tilde{\mathcal{E}}}(s_1) - \mathbb{E}_{\nu\sim(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}\left[V_\nu^{(i),\tilde{\mathcal{E}}}(s_1)\right]\right)\right| \leq 2\sqrt{SAH^3\mathbb{I}_k^{(\pi_{\mathrm{TS}}^{(i)}\times\pi^{k,(-i)})}(\tilde{\mathcal{E}};\mathcal{T}_{H+1}^k)}, \tag{H.8}$$

where we slightly modify the new environment $\hat{e}_k'$ in (G.15) such that its transition kernel is averaged w.r.t. $\tilde{\mathcal{E}}$ (instead of $\mathcal{E}$). Finally, we have

$$\mathrm{BR}_K^{\mathrm{NE}}(\pi) \leq \sum_{k=1}^K \mathbb{E}_{\mathcal{D}_k}\left[\sum_{i=1}^N \mathbb{E}_k\left(V_{\mu_{\mathcal{E}}^{(i),\pi^k}}^{(i),\tilde{\mathcal{E}}}(s_1) - V_{\pi^k}^{(i),\tilde{\mathcal{E}}}(s_1)\right)\right] + 2KN\epsilon \tag{H.9}$$

$$\leq \frac{SAH^3KN}{\lambda} + \lambda N\cdot\mathbb{I}^\pi(\tilde{\mathcal{E}};\mathcal{D}_{K+1}) + 2KN\epsilon \tag{H.10}$$

$$= 2N\sqrt{SAH^3K\cdot\mathbb{I}(\tilde{\mathcal{E}};\mathcal{D}_{K+1})} + 2KN\epsilon, \tag{H.11}$$

where the last step follows from choosing $\lambda = \sqrt{SAH^3K/\mathbb{I}(\tilde{\mathcal{E}};\mathcal{D}_{K+1})}$. We sometimes also express $\tilde{\mathcal{E}}$ as $\tilde{\mathcal{E}}_\epsilon$ to clearly indicate its correlation with the parameter ϵ. This completes the proof of Theorem 6.

Data availability

No data was used for the research described in the article.

References

[1] M. Brambilla, E. Ferrante, M. Birattari, M. Dorigo, Swarm robotics: a review from the swarm engineering perspective, Swarm Intell. 7 (2013) 1–41.
[2] S. Shalev-Shwartz, S. Shammah, A. Shashua, Safe, multi-agent, reinforcement learning for autonomous driving, arXiv preprint, arXiv:1610.03295, 2016.
[3] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.
[4] Y. Bai, C. Jin, T. Yu, Near-optimal reinforcement learning with self-play, Adv. Neural Inf. Process. Syst. 33 (2020) 2159–2170.
[5] C. Jin, Q. Liu, Y. Wang, T. Yu, V-learning—a simple, efficient, decentralized algorithm for multiagent reinforcement learning, Math. Oper. Res. (2023).
[6] B. Huang, J.D. Lee, Z. Wang, Z. Yang, Towards general function approximation in zero-sum Markov games, in: International Conference on Learning Representations, 2022.
[7] Q. Liu, T. Yu, Y. Bai, C. Jin, A sharp analysis of model-based reinforcement learning with self-play, in: International Conference on Machine Learning, 2021, pp. 7001–7010.
[8] W. Xiong, H. Zhong, C. Shi, C. Shen, T. Zhang, A self-play posterior sampling algorithm for zero-sum Markov games, in: International Conference on Machine Learning, 2022, pp. 24496–24523.
[9] S. Qiu, Z. Dai, H. Zhong, Z. Wang, Z. Yang, T. Zhang, Posterior sampling for competitive RL: function approximation and partial observation, in: Advances in Neural Information Processing Systems, 2023.
[10] D. Russo, B. Van Roy, Learning to optimize via information-directed sampling, in: Advances in Neural Information Processing Systems, 2014, pp. 1583–1591.
[11] D. Russo, B. Van Roy, Learning to optimize via information-directed sampling, Oper. Res. 66 (1) (2018) 230–252.
[12] B. Hao, T. Lattimore, Regret bounds for information-directed reinforcement learning, Adv. Neural Inf. Process. Syst. 35 (2022) 28575–28587.
[13] D.J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al., A tutorial on Thompson sampling, Found. Trends Mach. Learn. 11 (1) (2018) 1–96.
[14] N. Nikolov, J. Kirschner, F. Berkenkamp, A. Krause, Information-directed exploration for deep reinforcement learning, in: International Conference on Learning Representations, 2019.
[15] X. Lu, B. Van Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, Z. Wen, et al., Reinforcement learning, bit by bit, Found. Trends Mach. Learn. 16 (6) (2023) 733–865.
[16] J. Kirschner, A. Krause, Information directed sampling and bandits with heteroscedastic noise, in: Conference on Learning Theory, 2018, pp. 358–384.
[17] B. Hao, T. Lattimore, W. Deng, Information directed sampling for sparse linear bandits, in: Advances in Neural Information Processing Systems, 2021, pp. 16738–16750.
[18] J. Hao, T. Yang, H. Tang, C. Bai, J. Liu, Z. Meng, P. Liu, Z. Wang, Exploration in deep reinforcement learning: from single-agent to multiagent domain, IEEE Trans. Neural Netw. Learn. Syst. (2023).
[19] T. Berger, Rate-Distortion Theory, Wiley Encyclopedia of Telecommunications, 2003.
[20] D. Arumugam, B. Van Roy, Deciding what to learn: a rate-distortion approach, in: International Conference on Machine Learning, 2021, pp. 373–382.
[21] D. Arumugam, B. Van Roy, The value of information when deciding what to learn, Adv. Neural Inf. Process. Syst. 34 (2021) 9816–9827.
[22] D. Arumugam, B. Van Roy, Deciding what to model: value-equivalent sampling for reinforcement learning, Adv. Neural Inf. Process. Syst. 35 (2022) 9024–9044.
[23] C. Bai, L. Wang, J. Hao, Z. Yang, B. Zhao, Z. Wang, X. Li, Pessimistic value iteration for multi-task data sharing in offline reinforcement learning, Artif. Intell. 326 (2024) 104048.
[24] Y. Bai, C. Jin, Provable self-play algorithms for competitive reinforcement learning, in: International Conference on Machine Learning, 2020, pp. 551–560.
[25] Q. Xie, Y. Chen, Z. Wang, Z. Yang, Learning zero-sum simultaneous-move Markov games using function approximation and correlated equilibrium, in: Conference on Learning Theory, 2020, pp. 3674–3682.

[26] W. Mao, T. Başar, Provably efficient reinforcement learning in decentralized general-sum Markov games, Dyn. Games Appl. 13 (1) (2023) 165–186.
[27] Z. Song, S. Mei, Y. Bai, When can we learn general-sum Markov games with a large number of players sample-efficiently?, in: International Conference on Learning Representations, 2022.
[28] K. Zhang, Z. Yang, T. Başar, Multi-agent reinforcement learning: a selective overview of theories and algorithms, in: Handbook of Reinforcement Learning and Control, 2021, pp. 321–384.
[29] S. Qiu, X. Wei, J. Ye, Z. Wang, Z. Yang, Provably efficient fictitious play policy optimization for zero-sum Markov games with structured transitions, in: International Conference on Machine Learning, 2021, pp. 8715–8725.
[30] Z. Chen, D. Zhou, Q. Gu, Almost optimal algorithms for two-player zero-sum linear mixture Markov games, in: International Conference on Algorithmic Learning Theory, 2022, pp. 227–261.
[31] C. Jin, Q. Liu, T. Yu, The power of exploiter: provable multi-agent rl in large state spaces, in: International Conference on Machine Learning, 2022, pp. 10251–10279.
[32] Z. Liu, M. Lu, W. Xiong, H. Zhong, H. Hu, S. Zhang, S. Zheng, Z. Yang, Z. Wang, Maximize to explore: one objective function fusing estimation, planning, and exploration, in: Advances in Neural Information Processing Systems, 2023.
[33] C. Mao, Q. Zhang, Z. Wang, X. Li, On the role of general function approximation in offline reinforcement learning, in: The Twelfth International Conference on Learning Representations (ICLR), 2024.
[34] C. Jin, S. Kakade, A. Krishnamurthy, Q. Liu, Sample-efficient reinforcement learning of undercomplete pomdps, in: Advances in Neural Information Processing Systems, 2020, pp. 18530–18539.
[35] S. Hu, C.-w. Leung, H.-f. Leung, Modelling the dynamics of multiagent q-learning in repeated symmetric games: a mean field theoretic approach, Adv. Neural Inf. Process. Syst. 32 (2019).
[36] S. Hu, C.-W. Leung, H.-f. Leung, H. Soh, The dynamics of q-learning in population games: a physics-inspired continuity equation model, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, 2022, pp. 615–623.
[37] Q. Liu, C. Szepesvári, C. Jin, Sample-efficient reinforcement learning of partially observable Markov games, Adv. Neural Inf. Process. Syst. 35 (2022) 18296–18308.
[38] D. Foster, D.J. Foster, N. Golowich, A. Rakhlin, On the complexity of multi-agent decision making: from learning in games to partial monitoring, in: The Thirty Sixth Annual Conference on Learning Theory, 2023, pp. 2678–2792.
[39] S. Hu, H. Soh, G. Piliouras, The best of both worlds in network population games: reaching consensus and convergence to equilibrium, Adv. Neural Inf. Process. Syst. 36 (2023).
[40] Y. Wang, Q. Liu, Y. Bai, C. Jin, Breaking the curse of multiagency: provably efficient decentralized multi-agent RL with function approximation, in: Proceedings of Thirty Sixth Conference on Learning Theory, 2023.
[41] Q. Cui, K. Zhang, S. Du, Breaking the curse of multiagents in a large state space: RL in Markov games with independent linear function approximation, in: The Thirty Sixth Annual Conference on Learning Theory, 2023, pp. 2651–2652.
[42] N. Xiong, Z. Liu, Z. Wang, Z. Yang, Sample-efficient multi-agent rl: an optimization perspective, arXiv preprint, arXiv:2310.06243, 2023.
[43] F. Liu, S. Buccapatnam, N. Shroff, Information directed sampling for stochastic bandits with graph feedback, in: AAAI Conference on Artificial Intelligence, 2018.
[44] J. Kirschner, T. Lattimore, C. Vernade, C. Szepesvári, Asymptotically optimal information-directed sampling, in: Conference on Learning Theory, 2021, pp. 2777–2821.
[45] X. Lu, B. Van Roy, Information-theoretic confidence bounds for reinforcement learning, Adv. Neural Inf. Process. Syst. 33 (2019).
[46] A. Moradipari, M. Pedramfar, M.S. Zini, V. Aggarwal, Improved Bayesian regret bounds for Thompson sampling in reinforcement learning, in: Advances in Neural Information Processing Systems, 2023.
[47] S. Chakraborty, A.S. Bedi, A. Koppel, M. Wang, F. Huang, D. Manocha, Steering: Stein information directed exploration for model-based reinforcement learning, in: International Conference on Machine Learning, 2023, pp. 3949–3978.
[48] C. Grimm, A. Barreto, S. Singh, D. Silver, The value equivalence principle for model-based reinforcement learning, Adv. Neural Inf. Process. Syst. 33 (2020) 5541–5552.
[49] C. Grimm, A. Barreto, G. Farquhar, D. Silver, S. Singh, Proper value equivalence, Adv. Neural Inf. Process. Syst. 34 (2021) 7773–7786.
[50] D.J. Foster, S.M. Kakade, J. Qian, A. Rakhlin, The statistical complexity of interactive decision making, arXiv preprint, arXiv:2112.13487, 2021.
[51] Y. Xu, A. Zeevi, Bayesian design principles for frequentist sequential learning, in: International Conference on Machine Learning, 2023, pp. 38768–38800.
[52] T.M. Cover, Elements of Information Theory, John Wiley & Sons, 1999.
[53] J. Filar, K. Vrieze, Competitive Markov Decision Processes, Springer Science & Business Media, 2012.
[54] J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, Bayesian model averaging: a tutorial, Stat. Sci. 14 (4) (1999) 382–417.
[55] I. Osband, Z. Wen, S.M. Asghari, V. Dwaracherla, M. Ibrahimi, X. Lu, B. Van Roy, Epistemic neural networks, in: Advances in Neural Information Processing Systems, 2023.
[56] R. Blahut, Computation of channel capacity and rate-distortion functions, IEEE Trans. Inf. Theory 18 (4) (1972) 460–473.
[57] S. Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels, IEEE Trans. Inf. Theory 18 (1) (1972) 14–20.