Human-AI Collaboration Project

BY JACK YANSONG LI
University of Illinois Chicago
Email: yli340@uic.edu

1 Introduction (revised, please read)

Training an AI agent capable of cooperating with various types of humans stands as a central challenge in human-AI Interaction (HAI). This problem proves to be difficult because different humans can create varied environments for the AI agent to navigate. Additionally, the AI agent cannot presume rational behavior from humans within the collaboration setting YL: add citations here. While a significant number of studies in the field of multi-agent reinforcement learning have explored centralized settings [64][61][24], this approach can present challenges in situations where a central control for individual agents is neither practical nor feasible. It's critical to acknowledge that a significant body of work also explores decentralized settings. However, most of these studies necessitate some form of predetermined communication or coordination guidelines to address the constraints inherent in centralized frameworks. YL: add some citations of decentralized MARL here.

Consequently, collaborating with unknown humans without predetermined communication or coordination guidelines becomes important, giving rise to a research area named ad-hoc teamwork [40][7][50]. A pivotal subtask within the domain of ad-hoc teamwork is opponent modeling [6][5][12][58][29], a concept that, in the context of Human-AI (HAI) interactions, primarily entails modeling human behaviors and policies.

In this paper, we tackle a problem in the Human-AI (HAI) interaction domain where there are two participants: a human agent and an AI agent. The AI agent does not know the human agent's policy. However, the AI starts with a few initial guesses about the human's possible policies. The main goal of the AI is to find the best policy that can collaborate with human. This approach as falling within several known frameworks such as latent Markov Decision Processes (MDP) [35][27][15] and multi-task Reinforcement Learning (RL) [37] [54] [15].

To model the HAI problem, we use an episodic Markov decision process where the transitions and rewards are influenced by the policy that the human agent keeps secret. The AI agent begins with a set of initial guesses about the human's policy, grouped together in a finite hypothesis set $\mathcal{H}$. We discuss loosening the limitation of this finite set in a later Section 6.

We used the Maximize to Explore (MEX) algorithm mentioned in [38] to tackle the episodic MDP with an given hypothesis set. We confirm a sub-linear regret outcome in Section 5. Moreover, we found that the MEX algorithm can ignore human policies that are of the same type, allowing for a regret boundary that is smaller than the upper limit noted in [38]. The definition for policies be of the same type is introduced in Section 4. Furthermore, we applied the MEX algorithm with an infinite hypothesis set that encompasses the true policy. We demonstrated that utilizing MEX with a finite hypothesis set, which contains a policy nearly identical to the true policy in the infinite hypothesis set, can still achieve sub-linear regret that converges to a neighborhood of a optimal value. This aspect is elaborated in Section 6.

In our experiment, we developed a simplified environment of the Overcooked-AI [17], where agents are required to engage in a series of actions such as cooking, waiting, and delivering food. The simplified version of Overcooked-AI, focusing exclusively on the food delivery task. This simplification is essential as it enables us to focus clearly on the main challenges posed by the original environment but significantly reduce the size of both state and action spaces, leading to a considerable decrease in computational complexity.

We created the set of possible human policies using the best response dynamics method [52], where agents constantly modify their policies to best respond to the policies observed from other agents. The best response dynamics is guaranteed to converge to a Nash equilibrium policy, which equilibrium it reaches depends on the initial policy [41]. To increase the diversity of human policies, we hand-coded some policies as initial policies. Also, inspired by IPOMDP [25], we varied the total number of iterations in the best response dynamics to simulate humans exhibiting different levels of intelligence.

We compared our algorithm with Q-learning with UCB exploration [31][19], Upper confidence bound [36], optimistic posterior sampling [66], and UCRL2 [10] algorithms. Our results shows that (added experiment)

2 Related Work (haven't revise, do not read)

Previous research in the Human-AI (HAI) field tends to model human policy as a policy that closely aligns with the AI agent, with efforts to parameterize this closeness [42]. To evaluate these algorithms, several benchmark environments are available that aid in analyzing cooperative human-AI interaction tasks, including platforms like the two-player cooperative Atari game [57], bridge card game [40], and Overcooked-AI [16][54].

Human-AI Interaction: Previous research in the Human-AI (HAI) field model human policy as a policy that closely aligns with the AI agent, with efforts to parameterize this closeness [42]. Additionally, there are studies in Meta Reinforcement Learning (Meta RL) that work on deciphering the MDP the AI agent encounters, inherently learning the human agent's policy, since the structure of this MDP is influenced by the human agent's choices. To evaluate these algorithms, several benchmark environments are available that aid in analyzing cooperative human-AI interaction tasks, including platforms like the two-player cooperative Atari game [56], bridge card [39], and Overcooked-AI [17][52].

Human agent generation: Collecting human policies can be notably costly. Previous studies have developed methods for more efficient human policies generation. One such method utilizes an algorithm that identifies and selects policies based on a measure defined by the diversity of the best responses these policies can offer. The algorithm then maximizes this measure to find the policies that provide a diverse set of best responses [45]. Another strategy formulates human policies by running best response dynamics [52].

Ad-Hoc teamworks: Our work is closely related to ad-hoc teamworks [40][7][50] , especially the opponent modeling subtask. Barrett et. al. [12] introduces PLASTIC-Model and PLASTIC-Policy algorithms, the formal algorithm models the team-member by its transition dynamics and the latter models team-member by its policy. He et.al. [29] models the human agent's policy as a deep neural network.

YL: needs more time to read [6][5][58]

Partially observable Markov decision process (POMDP): The foundation of our problem is closely related to the partially observable Markov decision process [9][48], since each human policy in the hypothesis set can be viewed as a latent variable of the POMDP. The POMDP problem where we have latent variables are called latent MDPs (LMDP) [35]. LMDP has few different names, such as contextual decision process [27], multi-model MDP [49], multi-task RL [37][54][15], MOMDP/hidden model MDP [43][18][20][23], and concurrent MDP [16]. Beyond original POMDP, there are also some other settings that can cover our problem, such as interactive-POMDP [28][25], Augmented Bayes-Adaptive MDP (BAMDP) [57][22][26]. The model-based RL with UCB exploration algorithms [51][11] is also related to our setting.

MEX related algorithms: Our algorithm is based on MEX [38], in each episode, the algorithm chooses a human policy from the hypothesis set. On the other hand, the posterior sampling algorithms [55][47][66][2][65][3][4] updates a belief over the hypothesis set in each episode and draws a policy based on the current belief. Some methods like OLIVE [30] eliminates policy from current hypothesis set in each episode. Additionally, there is a method that trains one policy that is robust for all possible human policies in the hypothesis set [14].

MDP structure assumptions: Our regret analysis is based on the low generalized eluder coefficient assumption [66], which is a weaker assumption than low Eluder dimension/Bellman eluder dimension [44][32], low Bellman rank [30], Bellman completeness [62], Bilinear classes [21], and linear MDP structure [60][33]. The environment we used in the experiment is a tabular MDP which satisfies the low Bellman eluder dimension assumption [32]. Also, the regret analysis of infinite hypothesis set are related to agnostic online learning [13][46].

3 Prerequisite

In this work, we use an two-player episodic finite horizon MDP as the setting for human-AI collaboration, where the player 1 is the AI agent and the player 2 is the human agent. We assume that the different human types are captured by different policies and the human agent's real policy is fixed but unknown. We revised the MEX algorithm introduced in [38] so it can be used in our setting.

3.1 Episodic Finite Horizon MDP

We consider a two-player episodic finite horizon Markov game defined as $(S, A, \mathbb{P}, T, K, r, \gamma)$, where S is the joint state space and A is the action space for both players. In this paper, we only consider the tabular setting, i.e., $|S| < \infty$ and $|A| < \infty$. The transition kernel $\mathbb{P}$ and the expected reward r are defined as:

$$\mathbb{P}: S \times A \times A \rightarrow \Delta(S), \quad r: S \times A \times A \rightarrow [0, 1],$$

where $\Delta(S)$ is the probability simplex on joint state space. The time horizon and the total number of episodes are denoted by T and K. Also, we consider a discounted setting with a discount factor $\gamma < 1$. We use s_t^k to denote the joint state of episode k in time t and we use a_t^k and b_t^k to denote the action of episode k in time t for player 1 and player 2 respectively. The initial state is fixed for all episodes and denoted as s_1, i.e., $s_0^k = s_0$. The policy for player 1 is denoted as μ and defined as

$$\mu: S \times [T] \rightarrow \Delta(A),$$

where $[T] = \{0, 1, \ldots, T-1\}$. The policy for player 2 is denoted as π and defined as

$$\pi: S \times [T] \rightarrow \Delta(A).$$

The set of all policies available to player 1 is denoted by $\mathcal{U}$ and the set for player 2 is represented by Π. Now, we define the cumulative reward given policies (μ, π) as

$$V(\mu, \pi) = \mathbb{E}_{a_t \sim \mu(s_t, t), b_t \sim \pi(s_t, t)} \left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t, b_t) | \, s_0 = s_0, s_{t+1} \sim \mathbb{P}(s_t, a_t, b_t) \right].$$

In this study, we explore a scenario where player 2 adopts a fixed but unknown policy through all episodes. The true policy adopted for player 2 is denoted as π^*. The primary objective is to develop a series of policies $\{\mu^k\}_{k \in [K]}$ for player 1, which aim to reduce the regret defined by the equation

$$\mathrm{Reg}(K) = \sum_{k=1}^{K} \max_{\mu \in \mathcal{U}} V(\mu, \pi^*) - V(\mu^k, \pi^k).$$

Here, π^k is the guessed policy of player 2 in episode k (guessed by player 1). The guessed policies are selected from a pre-trained hypothesis set $\mathcal{H}$. This set contains potential behaviors of player 2. Some methods such as best response dynamics [41] can be employed to create $\mathcal{H}$. The detailed way to generate a hypothesis set will be discussed in Section (Experiment).

A significant challenge encountered in human-AI collaboration is the necessity for the AI agent to anticipate the actions of the human agent. In the context of our framework, the hypothesis set $\mathcal{H}$ embodies various potential human behaviors identified prior to the training phase, with the true policy, π^*, representing the actual behavior demonstrated by the human agent during the interaction with the AI agent.

3.2 Maximize to Explore Algorithm

TODO: introduce MEX in here, compare the original implementation and our implementation.

4 Classifying Different Types of Agents by Policies

In our previous discussion, we explained how the hypothesis set $\mathcal{H}$ captures the potential behaviors of player 2. However, $\mathcal{H}$ might include policies that result in similar optimal cumulative rewards. In instances like these, we should consider these policies as belonging to the same type. In this subsection, we will present a clear definition to classify the various types of policies. Building on these classifications, we will then develop a metric to assess the size of the hypothesis set $\mathcal{H}$, which will be smaller than the numerical count of $|\mathcal{H}|$.

4.1 Best Response Oracle

For each player 2's policy π, the set of all best response policies is denoted as $\mathrm{BR}(\pi)$, i.e.,

$$\mathrm{BR}(\pi) = \underset{\mu \in \mathcal{U}}{\mathrm{argmax}}\, V(\mu, \pi).$$

For any player 2's policy π, we assume the existence of an oracle that can return a best response from $\mathrm{BR}(\pi)$.

Definition 1. ***(Oracle)*** *A best response oracle ψ refers to a function that, upon receiving policies as input, yields a best response as its output, i.e., ψ is a function $\psi: \mathcal{H} \rightarrow \mathcal{U}$ such that*

$$\psi(\pi) \in \mathrm{BR}(\pi).$$

With the definition of an oracle, we can categorize policies within a hypothesis set into various types.

4.2 Type Number

Definition 2. ***(ψ-type)*** *We call two policies π and π' to be of the same type under oracle ψ if we have*

$$V(\psi(\pi), \pi) = V(\psi(\pi'), \pi').$$

The relationship is denoted as $\pi \overset{\psi}{\sim} \pi'$. On the contrary, two policies π and π' not of the same type under oracle ψ is denoted as $\pi \overset{\psi}{\nsim} \pi'$.

Definition 3. *We call a set of policies Π be type-independent under oracle ψ if for all $\pi \in \Pi$ and $\pi' \in \Pi$ such that $\pi \neq \pi'$, we have $\pi \overset{\psi}{\nsim} \pi'$.*

The ψ-type characterization gives rise to a measurement of quantity for the set of policies $\mathcal{H}$, denoted by $n^{\psi}(\mathcal{H})$.

Definition 4. *Given a hypothesis set $\mathcal{H}$, the type number $n^{\psi}(\mathcal{H})$ under oracle ψ is defined as the size of a largest type-independent subset of $\mathcal{H}$, i.e.,*

$$n^{\psi}(\mathcal{H}) = \max\{|\Pi|: \Pi \subset \mathcal{H}, \Pi \textit{ is type-independent under oracle } \psi\}.$$

It is easy to verify that $n^{\psi}(\mathcal{H}) \le |\mathcal{H}|$. In the next section, we will show that the regret of the MEX algorithm depends on $n^{\psi}(\mathcal{H})$ instead of $|\mathcal{H}|$.

5 Regret Analysis for Finite Hypothesis Set

In this section, we restrict our discussion to cases where the cardinality of the hypothesis set is finite, i.e., $|\mathcal{H}| < \infty$. This condition is emphasized through the notation $\mathcal{H}_{\text{fin}}$. We also assume that the realization assumption holds, i.e., $\pi^* \in \mathcal{H}_{\text{fin}}$.

5.1 Generalized Eluder Coefficient

Our sub-linear regret result based on the following assumption.

Assumption 5. *(Low generalized eluder coefficient [38][66])* Given $\varepsilon > 0$, there exists $d(\varepsilon) > 0$, such that for any $\{\pi^k\}_{k\in[K]} \subset \mathcal{H}$, $\{\psi(\pi^k)\}_{k\in[K]} \subset \mathcal{U}$,

$$\sum_{k=1}^{K} V(\mu^k, \pi^k) - V(\mu^k, \pi^*) \le \inf_{\mu>0} \left\{ \frac{\mu}{2} \sum_{k=1,h=1}^{K,H} \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \mu^s}[\ell(\pi^k; \xi_h)] + \varphi(\mu, \varepsilon, H, K) \right\}, \tag{1}$$

where $\varphi(\mu, \varepsilon, H, K) = d(\varepsilon)/(2\mu) + \sqrt{d(\varepsilon)HK} + \varepsilon HK$. The player 1's policy in episode s is given by $\mu^s = \psi(\pi^s)$. The discrepancy function ℓ_{π^s} is the *Hellinger distance*. Given data $\xi_h = (s_h, a_h, r_h, s_{h+1})$, we define

$$\ell(\pi^k, \xi_h) = D_H(\mathbb{P}(\cdot|\, s_h, a_h, \pi^k(s_h)) ||\, \mathbb{P}(\cdot|\, s_h, a_h, \pi^*(s_h))),$$

where $D_H(\cdot||\cdot)$ denotes the Hellinger distance. We denote the smallest number $d(\varepsilon) > 0$ satisfying condition (1) as $d_{\text{GEC}}(\varepsilon)$.

Intuitively, the low generalized eluder coefficient assumption states that, in the long run, if the hypothesis $\{\pi^k\}_{k\in[K]}$ has a small in-sample training error, i.e., the term

$$\mathbb{E}_{\xi_h \sim \mu^s}[\ell(\pi^k; \xi_h)]$$

is small, then, the prediction error $V(\mu^k, \pi^k) - V(\mu^k, \pi^*)$ will also be small. In [30], they showed that a tabular MDP has a low Bellman rank, which implies a low Bellman eluder dimension [32]. In [66], they showed that any MDP that satisfies the low Bellman eluder dimension condition will have a low GEC condition, which implies that our setting satisfies condition (1).

5.2 Main Theorem: Finite Hypothesis Set

Theorem 6. *Given an MDP with generalized eluder coefficient $d_{\text{GEC}}(\cdot)$ and a finite hypothesis class $\mathcal{H}_{\text{fin}}$ with $\pi^* \in \mathcal{H}_{\text{fin}}$, by setting*

$$\eta = \sqrt{\frac{d_{\text{GEC}}(1/\sqrt{HK})}{\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta) \cdot HK}},$$

the regret of the MEX algorithm applying on $\mathcal{H}_{\text{fin}}$ with oracle ψ after K episodes is upper bounded by, with probability at least $1-\delta$,

$$\text{Regret}(K) \lesssim \sqrt{d_{\text{GEC}}(1/\sqrt{HK}) \cdot \log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta) \cdot HK}.$$

Proof. See Appendix 8.2 □

The sole term related to the size of the hypothesis set is $n^{\psi}(\mathcal{H}_{\rm fin})$. Consequently, the magnitude of regret is solely influenced by the type number associated with a hypothesis set, as opposed to the cardinality of the hypothesis set. This phenomenon occurs because policies that are categorized under the same type by policy ψ yield identical rewards when implemented in the MEX algorithm.

The type number $n^{\psi}(\mathcal{H}_{\rm fin})$ depends on the choice of the oracle ψ, which makes it hard to verify when the explicit form of ψ is not given. However, we can introduce a stronger notion of type and verify the upper bound of $n^{\psi}(\mathcal{H}_{\rm fin})$. This stronger notion of type does not depends on the choice of oracle ψ.

Definition 7. *(Strong) We call two policies π and π' to be of the same s-type if*

$$V(\mu,\pi)=V(\mu,\pi')=V(\mu',\pi)=V(\mu',\pi')$$

for all $\mu\in\mathrm{BR}(\pi)$ and $\mu'\in\mathrm{BR}(\pi')$. The relationship are denoted as $\pi\overset{s}{\sim}\pi'$.

Similar to the definition of type number under oracle ψ, we can define strong type number $n_{\rm stype}(\mathcal{H})$.

Lemma 8. *$n^{\psi}(\mathcal{H})\le n_{\rm stype}(\mathcal{H})$ for all ψ be a best response oracle.*

Proof. By definition. □

Example 9. For any two distinct policies, π and π', if they both have the same unique best response μ, we can deduct that $\pi\overset{s}{\sim}\pi'$ by definition. To illustrate this in a real-world context, let's consider a scenario involving two cars traveling on a road with three lanes. The AI-operated car is in the left lane, while the human-driven car occupies the middle lane. Suppose the AI car intends to accelerate to move ahead of the human car. In response, the human has three options: move to the right, shift to the left, or maintain their current position. If the human's policy tends towards moving right or staying put, the AI car will consistently choose to speed up, given that both these responses from the human elicit the same best response from the AI. Consequently, these two human policies can be considered to be of the same strong type, or s-type.

In the next section, we will generalize the regret analysis to an infinite hypothesis set.

6 Regret Analysis for Infinite Hypothesis Set

In this subsection, we discuss the cases where the cardinality of the hypothesis set is infinite, i.e., $|\mathcal{H}|=\infty$. This condition is emphasized through the notation $\mathcal{H}_{\rm inf}$. We keep assume that the realization assumption holds, i.e., $\pi^*\in\mathcal{H}_{\rm inf}$.

6.1 Approximate an Infinite Hypothesis Set by a Finite Hypothesis Set

A direct approach to handling an infinite hypothesis set is to approximate it by a finite hypothesis set. First, we outline what makes a good approximation. Let's first define the following quantity,

$$V^*(\pi)\triangleq\max_{\mu\in\mathcal{U}}V(\mu,\pi)=V(\psi(\pi),\pi)$$

The quantity V^* serves as a value function for player 2's policy given player 2 will return a best response. In the tabular case, we can regard each policy π as a vector in $\mathbb{R}^{|S||A|T}$. Now given a norm $\|\cdot\|$ such as L_2-norm defined on Π, we have the following definition,

Definition 10. *(ε-approximation) A finite hypothesis set $\mathcal{H}_{\rm fin}$ is called an ε-optimal approximation of $\mathcal{H}_{\rm inf}$ if there exist a $\pi\in\mathcal{H}_{\rm fin}$ such that $\|\pi-\pi^*\|\le\varepsilon$.*

Based on this definition, we make the following Assumption.

Assumption 11. The value function for player 2's policy $V^*(\pi)$ is Lipschitz continuous, i.e., if $\|\pi-\pi^*\|\leq\varepsilon$, we have

$$|V^*(\pi)-V^*(\pi^*)|\leq L_V\varepsilon,$$

for some constant $L_V>0$.

The assumption holds for MDPs where the transition kernel and the reward function are both Lipschitz continuous, see Proposition 2 in [53]. Similar results can also be found in [8][34][63][59][1]

Example 12. Many works (TODO: add citations) in ad-hoc teamworks assumes an finite hypothesis set $\mathcal{H}_{\text{fin}}$ to capture all potential policies adopted by teammates. However, the real policy the teammates adopts might deviate from the given hypothesis set $\mathcal{H}_{\text{fin}}$ due to reasons like bounded rationality. Thus, the real hypothesis set that captures the deviation is defined as

$$\mathcal{H}_{\text{inf}}=\{\pi|\,\|\pi-\pi'\|\leq\varepsilon,\pi'\in\mathcal{H}_{\text{fin}}\}.$$

Thus, the given hypothesis set $\mathcal{H}_{\text{fin}}$ serves as an ε-optimal approximation of the real hypothesis set $\mathcal{H}_{\text{inf}}$

Example 13. Given a specific parameterization $\mathcal{N}$, we define $\mathcal{H}_{\text{inf}}$ as the set comprising all policies characterized by the set all possible parameters Θ that is in accordance with the specified structure, formally represented as,

$$\mathcal{H}_{\text{inf}}=\{\pi|\,\pi\in\mathcal{N}(\theta),\theta\in\Theta\}.$$

We proceed to create a discretization of Θ, denoted as $\hat{\Theta}$. The finite approximation set $\mathcal{H}_{\text{fin}}$ is defined as

$$\mathcal{H}_{\text{fin}}=\{\pi|\,\pi\in\mathcal{N}(\theta),\theta\in\hat{\Theta}\}.$$

By the choice of discretization interval, we can ensure that $\|\pi-\pi^*\|\leq\varepsilon$.

6.2 Main Theorem: Infinite Hypothesis Set

Now, given an infinite hypothesis set $\mathcal{H}_{\text{inf}}$ with an ε-optimal approximation set $\mathcal{H}_{\text{fin}}$, we are prepared to execute the MEX algorithm within the confines of $\mathcal{H}_{\text{fin}}$. The regret analysis is given in the following Theorem. Before proof the theorem, we denote π^*_{det} as ε-approximate policy w.r.t. π^* defined as

$$\pi^*_{\text{det}}\triangleq\underset{\pi\in\mathcal{H}}{\operatorname{argmin}}\,|V^*(\pi)-V^*(\pi^*)|.$$

Theorem 14. *Given an MDP with generalized eluder coefficient $d_{\text{GEC}}(\cdot)$ and an infinite hypothesis class $\mathcal{H}_{\text{inf}}$ with $\pi^*\in\mathcal{H}_{\text{inf}}$. For any ε-optimal approximation $\mathcal{H}_{\text{fin}}$ of $\mathcal{H}_{\text{inf}}$, by setting*

$$\eta=\sqrt{\frac{d_{\text{GEC}}(1/\sqrt{HK})}{\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta)\cdot HK}},$$

the regret of the MEX algorithm applying on $\mathcal{H}_{\text{fin}}$ with oracle ψ after K episodes is upper bounded by, with probability at least $1-\delta$,

$$\text{Regret}(K)\lesssim\sqrt{d_{\text{GEC}}(1/\sqrt{HK})\cdot\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta)\cdot HK}+KL_V\varepsilon.$$

Proof. By the choice of π^k, we have

$$V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) - \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^*_{\text{det}}) \leq V(\mu^k, \pi^k) - \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^k)$$

for all $k \in [K]$. By Definition 10,

$$V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) \geq V(\psi(\pi^*), \pi^*) - L_V \varepsilon.$$

Thus,

$$V(\psi(\pi^*), \pi^*) - V(\mu^k, \pi^k) \leq \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^*_{\text{det}}) - \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^k) + L_V \varepsilon.$$

Follow the same procedure in the proof of Theorem 6 leads to the proof. $\square$

Remark 15. The linear term $K L_V \varepsilon$ cannot be eliminated. Consider the best case where $\pi^k = \pi^*_{\text{det}}$ for all $k \in [K]$. The regret is

$$\begin{aligned} \text{Regret}(K) &= \sum_{k=1}^{K} V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) - V(\psi(\pi^*), \pi^*) \\ &= K(V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) - V(\psi(\pi^*), \pi^*)) \\ &\leq K L_V \varepsilon. \end{aligned}$$

7 Numerical Experiments

8 Appendix

8.1 Lemmas

Lemma 16. *With probability at least $1-\delta$, for any $(h,k) \in [H] \times [K]$, $\mu^s \in \text{BR}(\pi^s)$, and $\pi \in \Pi$*

$$L_h^{k-1}(\pi^*) - L_h^{k-1}(\pi) \leq -2 \sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \mu^s}[\ell_{\pi^s}(\pi; \xi_h)] + 2\log(H|\Pi|/\delta).$$

Proof. Given $\pi \in \mathcal{H}$, we denote the random variable $X_{h,\pi}^k$ as

$$X_{h,\pi}^k = \log\left(\frac{\mathbb{P}_{h,\pi^*}(s_{h+1}^k | s_h^k, a_h^k)}{\mathbb{P}_{h,\pi}(s_{h+1}^k | s_h^k, a_h^k)}\right).$$

Now we define a filtration $\{\mathcal{F}_{h,k}\}_{k=1}^K$ as (B.25) in [38]. Thus we have $X_{h,\pi}^k \in \mathcal{F}_{h,k}$. Therefore, by applying Lemma D.1 in [38], we have that with probability at least $1-\delta$, for any $(h,k) \in [H] \times [K]$, and $\pi \in \Pi$, we have

$$-\frac{1}{2} \sum_{s=1}^{k-1} X_{h,\pi}^s \leq \sum_{s=1}^{k-1} \log \mathbb{E}\left[\exp\left\{-\frac{1}{2} X_{h,\pi}^s\right\} | \mathcal{F}_{h,s-1}\right] + \log(H|\Pi|/\delta). \tag{2}$$

Meanwhile, by (B.27) in [38], for any $\mu^s \in \mathrm{BR}(\pi^s)$, the conditional expectation equals to

$$\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,\pi}^s\right\} \mid \mathcal{F}_{h,s-1}\right] = 1 - \mathbb{E}_{(s_h^s, a_h^s)\sim\mu^s}[D_H(\mathbb{P}_{h,\pi^*}(\cdot \mid s_h^s, a_h^s) || \mathbb{P}_{h,\pi}(\cdot \mid s_h^s, a_h^s))]. \tag{3}$$

Denote $\mathbb{E}_{(s_h^s, a_h^s)\sim\mu^s}[D_H(\mathbb{P}_{h,\pi^*}(\cdot \mid s_h^s, a_h^s) || \mathbb{P}_{h,\pi^s}(\cdot \mid s_h^s, a_h^s))]$ as $\mathbb{E}_{\xi_h\sim\mu^s}[\ell_{\pi^s}(\pi;\xi_h)]$. Using the fact $\log(x) \leq x - 1$ and substituting (3) into (2) finishes the proof. □

Lemma 17. *If $a = \operatorname{argmax}_{x\in\mathcal{X}} [f(x) + g(x)]$ and $b = \operatorname{argmax}_{x\in\mathcal{X}} f(x)$, then, $f(b) \geq f(a)$ and*

$$f(b) - f(a) \geq g(b) - g(a).$$

Proof. By definition,

$$f(a) + g(a) \geq f(x) + g(x)$$

for all $x \in \mathcal{X}$. Let $x = b$, we have

$$f(a) + g(a) \geq f(b) + g(b).$$

Rearranging the above formula gives us

$$f(b) - f(a) \geq g(b) - g(a).$$

Similarly, by definition

$$f(b) \geq f(x)$$

for all $x \in \mathcal{X}$. Let $x = a$ gives $f(b) \geq f(a)$ □

8.2 Proof of Theorem 6

Proof. We decompose the regret into two terms,

$$\begin{aligned}
\mathrm{Regret}(K) &\triangleq \sum_{k=1}^{K} V(\psi(\pi^*), \pi^*) - V(\psi(\pi^k), \pi^*) \\
&= \underbrace{\sum_{k=1}^{K} V(\psi(\pi^*), \pi^*) - V(\psi(\pi^k), \pi^k)}_{\text{Term (i)}} + \underbrace{\sum_{k=1}^{K} V(\psi(\pi^k), \pi^k) - V(\psi(\pi^k), \pi^*)}_{\text{Term (ii)}}.
\end{aligned}$$

Term (i). By the choice of π^k, we have

$$V(\psi(\pi^*), \pi^*) - \eta\sum_{h=1}^{H} L_h^{k-1}(\pi^*) \leq V(\mu^k, \pi^k) - \eta\sum_{h=1}^{H} L_h^{k-1}(\pi^k)$$

for all $k \in [K]$. Thus,

$$V(\psi(\pi^*), \pi^*) - V(\mu^k, \pi^k) \leq \eta\sum_{h=1}^{H} L_h^{k-1}(\pi^*) - \eta\sum_{h=1}^{H} L_h^{k-1}(\pi^k). \tag{4}$$

for any $\pi^k \overset{\psi}{\sim} \pi^{k'}$, we have

$$V(\psi(\pi^k), \pi^k) = V(\psi(\pi^{k'}), \pi^{k'}).$$

Thus, an upper bound for $V(\psi(\pi^*), \pi^*) - V(\mu^k, \pi^k)$ is also an upper bound for $V(\psi(\pi^*), \pi^*) - V(\mu^{k'}, \pi^{k'})$. Applying Lemma 16, we have that with probability at least $1-\delta$, for any $(h,k) \in [H] \times [K]$, $\mu^s = \psi(\pi^s)$ and $\pi^k \in \mathcal{H}_{\text{fin}}$,

$$L_h^{k-1}(\pi^*) - L_h^{k-1}(\pi^k) \leq -2\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \mu^s}[\ell_{\pi^s}(\pi; \xi_h)] + 2\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta).$$

Substituting the above equation into (4) gives us that with probability at least $1-\delta$, for any $k \in [K]$, $\mu^s = \psi(\pi^s)$ and $\pi^k \in \mathcal{H}_{\text{fin}}$

$$V(\psi(\pi^*), \pi^*) - V(\mu^k, \pi^k) \leq -2\eta\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \mu^s}[\ell_{\pi^s}(\pi; \xi_h)] + 2H\eta\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta).$$

Summing over $[K]$ gives us

$$\text{Term (i)} \leq -2\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \mu^s}[\ell_{\pi^s}(\pi; \xi_h)] + 2\eta KH\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta).$$

Term (ii). Follow the proof of Theorem 4.4 in [38], we have that for all $\mu^s = \psi(\pi^s)$

$$\text{Term (ii)} \leq 2\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h \sim \mu^s}[\ell_{\pi^s}(\pi; \xi_h)] + \frac{d_{\text{GEC}}(\varepsilon_{\text{conf}})}{8\eta} + \sqrt{d_{\text{GEC}}(\varepsilon_{\text{conf}})HK} + \varepsilon_{\text{conf}}HK.$$

Combining Term (i) and Term (ii).

$$\begin{aligned}\text{Regret}(K) &= \text{Term (i)} + \text{Term (ii)}\\ &\leq 2\eta KH\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta) + \frac{d_{\text{GEC}}(\varepsilon_{\text{conf}})}{8\eta} + \sqrt{d_{\text{GEC}}(\varepsilon_{\text{conf}})HK} + \varepsilon_{\text{conf}}HK.\end{aligned}$$

Set $\varepsilon_{\text{conf}} = 1/\sqrt{HK}$ and

$$\eta = \sqrt{\frac{d_{\text{GEC}}(1/\sqrt{HK})}{\log(Hn^{\psi}(\mathcal{H}_{\text{fin}})/\delta) \cdot HK}}$$

leads to the proof. □

9 Do Not Read The Following

An optimal policy is denoted as (μ^*, π^*) and defined by

$$(\mu^*, \pi^*) \in \underset{(\mu,\pi)\in\mathcal{U}\times\Pi}{\operatorname{argmax}} V(\mu, \pi).$$

Bibliography

[1] Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 64-66. PMLR, jul 2020.

[2] Alekh Agarwal and Tong Zhang. Model-based RL with Optimistic Posterior Sampling: Structural Conditions and Sample Complexity. oct 2022.

[3] Alekh Agarwal and Tong Zhang. Non-Linear Reinforcement Learning in Large Action Spaces: Structural Conditions and Sample-efficiency of Posterior Sampling. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2776-2814. PMLR, jun 2022.

[4] Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. mar 2020.

[5] Stefano V. Albrecht, Jacob W. Crandall, and Subramanian Ramamoorthy. Belief and Truth in Hypothesised Behaviours. *Artificial Intelligence*, 235:63-94, jun 2016.

[6] Stefano V. Albrecht and Subramanian Ramamoorthy. A Game-Theoretic Model and Best-Response Learning Method for Ad Hoc Coordination in Multiagent Systems. jun 2015.

[7] Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66-95, may 2018.

[8] Kavosh Asadi, Dipendra Misra, and Michael L. Littman. Lipschitz Continuity in Model-based Reinforcement Learning. jul 2018.

[9] K.J Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174-205, feb 1965.

[10] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[11] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. jul 2017.

[12] Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132-171, jan 2017.

[13] Shai Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic Online Learning. In *Annual Conference Computational Learning Theory*. 2009.

[14] Luca F. Bertuccelli, Albert Wu, and Jonathan P. How. Robust Adaptive Markov Decision Processes: Planning with Model Uncertainty. *IEEE Control Systems Magazine*, 32(5):96-109, oct 2012.

[15] Emma Brunskill and Lihong Li. Sample Complexity of Multi-task Reinforcement Learning. sep 2013.

[16] Peter Buchholz and Dimitri Scheftelowitsch. Computation of weighted sums of rewards for concurrent MDPs. *Math Meth Oper Res*, 89(1):1-42, feb 2019.

[17] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the Utility of Learning about Humans for Human-AI Coordination. Oct 2019.

[18] Iadine Chades, Josie Carwardine, Tara Martin, Samuel Nicol, Regis Sabbadin, and Olivier Buffet. MOMDPs: A Solution for Modelling Adaptive Management Problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):267-273, 2012.

[19] Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB Exploration is Sample Efficient for Infinite-Horizon MDP. sep 2019.

[20] Finale Doshi-Velez and George Konidaris. Hidden Parameter Markov Decision Processes: A Semiparametric Regression Approach for Discovering Latent Task Parametrizations. aug 2013.

[21] Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear Classes: A Structural Framework for Provable Generalization in RL. jul 2021.

[22] Michael O'Gordon Duff. Optimal learning: Computational procedures for Bayes -adaptive Markov decision processes. *Doctoral Dissertations Available from Proquest*, pages 1-247, jan 2002.

[23] A. Fern, S. Natarajan, K. Judah, and P. Tadepalli. A Decision-Theoretic Model of Assistance. *Journal of Artificial Intelligence Research*, 50:71-104, may 2014.

[24] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual Multi-Agent Policy Gradients. dec 2017.

[25] P. J. Gmytrasiewicz and P. Doshi. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24:49-79, jul 2005.

[26] Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1025-1033. Red Hook, NY, USA, dec 2012. Curran Associates Inc.

[27] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov Decision Processes. feb 2015.

[28] Yanlin Han and Piotr Gmytrasiewicz. Learning Others' Intentional Models in Multi-Agent Settings Using Interactive POMDPs. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[29] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent Modeling in Deep Reinforcement Learning. sep 2016.

[30] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable. dec 2016.

[31] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning Provably Efficient? jul 2018.

[32] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. jul 2021.

[33] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 2137-2143. PMLR, jul 2020.

[34] Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic Analysis of Biased Stochastic Approximation Scheme. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1944-1974. PMLR, jun 2019.

[35] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. RL for Latent MDPs: Regret Guarantees and a Lower Bound. feb 2021.

[36] T. L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4-22, mar 1985.

[37] Yao Liu, Zhaohan Guo, and Emma Brunskill. PAC Continuous State Online Multitask Reinforcement Learning with Identification. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pages 438-446. Richland, SC, may 2016. International Foundation for Autonomous Agents and Multiagent Systems.

[38] Zhihan Liu, Miao Lu, Wei Xiong, Han Zhong, Hao Hu, Shenao Zhang, Sirui Zheng, Zhuoran Yang, and Zhaoran Wang. One Objective to Rule Them All: A Maximization Objective Fusing Estimation and Planning for Exploration. may 2023.

[39] Edward Lockhart, Neil Burch, Nolan Bard, Sebastian Borgeaud, Tom Eccles, Lucas Smaira, and Ray Smith. Human-Agent Cooperation in Bridge Bidding. nov 2020.

[40] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. A Survey of Ad Hoc Teamwork Research. In Dorothea Baumeister and Jörg Rothe, editors, *Multi-Agent Systems*, volume 13442, pages 275-293. Springer International Publishing, Cham, 2022.

[41] Dov Monderer and Lloyd S. Shapley. Potential Games. *Games and Economic Behavior*, 14(1):124-143, may 1996.

[42] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618-634, jun 2017.

[43] Sylvie C. W. Ong, Shao Wei Png, David Hsu, and Wee Sun Lee. Planning under Uncertainty for Robotic Tasks with Mixed Observability. *Int. J. Rob. Res.*, 29(8):1053-1068, jul 2010.

[44] Ian Osband and Benjamin Van Roy. Model-based Reinforcement Learning and the Eluder Dimension. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[45] Arrasy Rahman, Elliot Fosong, Ignacio Carlucho, and Stefano V. Albrecht. Generating Teammates for Training Robust Ad Hoc Teamwork Agents via Best-Response Diversity. *Transactions on Machine Learning Research*, may 2023.

[46] Stephane Ross and J. Andrew Bagnell. Agnostic System Identification for Model-Based Reinforcement Learning. *ArXiv:1203.1007 [cs, stat]*, jul 2012.

[47] Daniel Russo and Benjamin Van Roy. Learning to Optimize Via Posterior Sampling. feb 2014.

[48] Richard D. Smallwood and Edward J. Sondik. The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon. *Operations Research*, 21(5):1071-1088, 1973.

[49] Lauren N. Steimle, David L. Kaufman, and Brian T. Denton. Multi-model Markov decision processes. *IISE Transactions*, pages 1-16, may 2021.

[50] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. *AAAI*, 24(1):1504-1509, jul 2010.

[51] Alexander L. Strehl and Michael L. Littman. An analysis of model-based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309-1331, dec 2008.

[52] D. J. Strouse, Kevin R. McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with Humans without Human Data. jan 2022.

[53] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[54] Matthew E. Taylor and Peter Stone. Transfer Learning for Reinforcement Learning Domains: A Survey. *J. Mach. Learn. Res.*, 10:1633-1685, dec 2009.

[55] William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285-294, 1933.

[56] Paul Tylkin, Goran Radanovic, and David C Parkes. Learning Robust Helpful Behaviors in Two-Player Cooperative Atari Environments. 2021.

[57] D. J. White. Bayesian Decision Problems and Markov Chains. *Royal Statistical Society. Journal. Series A: General*, 132(1):106-107, jan 1969.

[58] Annie Xie, Dylan P. Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. Learning Latent Representations to Influence Multi-Agent Interaction. nov 2020.

[59] Pan Xu, Felicia Gao, and Quanquan Gu. Sample Efficient Policy Gradient Methods with Recursive Variance Reduction. aug 2021.

[60] Lin Yang and Mengdi Wang. Sample-Optimal Parametric Q-Learning Using Linearly Additive Features. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6995-7004. PMLR, may 2019.

[61] Xinghu Yao, Chao Wen, Yuhui Wang, and Xiaoyang Tan. SMIX($\ lambda$): Enhancing Centralized Value Functions for Cooperative Multi-Agent Reinforcement Learning. aug 2020.

[62] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning Near Optimal Policies with Low Inherent Bellman Error. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10978-10989. PMLR, nov 2020.

[63] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies. *SIAM J. Control Optim.*, 58(6):3586-3612, jan 2020.

[64] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. apr 2021.

[65] Tong Zhang. Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning. oct 2021.

[66] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond. jun 2023.