POSTERIOR SAMPLING FOR MULTI-AGENT REINFORCEMENT LEARNING: SOLVING EXTENSIVE GAMES WITH IMPERFECT INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Posterior sampling for reinforcement learning (PSRL) is a useful framework for making decisions in an unknown environment. PSRL maintains a posterior distribution of the environment and then makes planning on the environment sampled from the posterior distribution. Though PSRL works well on single-agent reinforcement learning problems, how to apply PSRL to multi-agent reinforcement learning problems is relatively unexplored. In this work, we extend PSRL to two-player zero-sum extensive-games with imperfect information (TZIEG), which is a class of multi-agent systems. More specifically, we combine PSRL with counterfactual regret minimization (CFR), which is the leading algorithm for TZIEG with a known environment. Our main contribution is a novel design of interaction strategies. With our interaction strategies, our algorithm provably converges to the Nash Equilibrium at a rate of $O(\sqrt{\log T/T})$. Empirical results show that our algorithm works well.

1 INTRODUCTION

Reinforcement Learning (RL) (Sutton & Barto, 2018) provides a framework for decision-making problems in an unknown environment, such as robotics control. In an RL problem, agents improve their strategies by gaining information from iteratively interacting with the environment. One of the key challenges in RL is how to interact with the environment.

Posterior sampling for RL (PSRL) (Strens, 2000) provides a useful framework for deciding how to interact with the environment. PSRL maintains a posterior distribution for the underlying environment and uses an environment sampled from this posterior to compute its interaction strategies. The interaction strategies are then used to interact with the environment to collect data. The design of the interaction strategies relies on specific problems. For example, in a single-agent RL (SARL) problem, PSRL takes the strategy with the maximum expected reward on the sampled environment as the interaction strategy (Osband et al., 2013).Theoretical and empirical results (Osband & Van Roy, 2016) both demonstrate that PSRL is one of the near-optimal methods for SARL. Although PSRL is a Bayesian-style algorithm, empirical evaluation (Chapelle & Li, 2011) and theoretical analysis on the multi-armed bandit problems (Agrawal & Goyal, 2017) suggest that it also enjoys good performance in a non-Bayesian setting.

However, applying PSRL to multi-agent RL (MARL) requires additional design on the interaction strategies. This is because the goal of MARL is quite different from that of SARL. In an MARL problem, each agent still aims to maximize its own reward, but the reward of an agent relies not only on the environment, but also on the strategies of other agents. Therefore, in MARL, the goal of learning is generally referred to finding a Nash Equilibrium (NE) where no agent is willing to deviate its strategy individually. So we should design the interaction strategies with which the agents can find or approximate the NE efficiently.

More specifically, we consider the RL problem in extensive games (Osborne & Rubinstein, 1994). Extensive games provide a unified model for sequential decision-making problems in which agents take actions in turn. In particular, we concentrate on two-player zero-sum imperfect information games (TZIEG) where there are two players gaining opposite rewards and a chance player to model the transition of the environment. Imperfect information here means that agents can keep their

own private information, such as the private pokers in poker games. Games with imperfect information are also fundamental to many practical issues such as economics and security. When the environment (i.e. the transition functions of the chance player and the reward functions) is known, counterfactual regret minimization (CFR) (Zinkevich et al., 2008) is the leading algorithm in approximating the NE in a TZIEG. However, in the RL setting where the environment is unknown, CFR is not applicable.

In this work, we present a posterior sampling algorithm for TZIEGs with the technique of CFR. That is, we apply CFR to the environment sampled from the posterior distribution. Our main contribution is to propose a novel design of interaction strategies for the RL problem of TZIEGs. With the proposed strategies, we show that our algorithm can provably converge to an approximate NE at a rate of $O(\sqrt{\log T/T})$. Empirical results show that our algorithm works well.

2 PRELIMINARY

In this section, we formulate the problem of TZIEGs and then we introduce the framework of the posterior sampling for reinforcement learning. Finally we briefly introduce the counterfactual regret minimization.

2.1 PROBLEM FORMULATION

We now formulate the problem. Firstly, we present the definition of extensive games in Defn. 1 (See (Osborne & Rubinstein, 1994, pg. 200) for a formal definition.).

Definition 1 (Extensive game)**.** *An extensive game can be described by a game tree, H. Each node in H is a history which is a sequence of past actions. Suppose there are N players participating in the game and let $\mathcal{C}$ denote the chance player. Let $[N] = \{1, 2, \cdots, N\}$ and $P(h) \in [N] \cup \{\mathcal{C}\}$ denote the player who is going to take an action at h. The game starts at the root of H, i.e., the empty history. At each non-terminal history, there is one player taking an action, and then the game moves to the next history accordingly. At a terminal history h, player i receives a reward sampled from a distribution $r^{*,i}(h)$. We assume the support of $r^{*,i}(h)$ is in $[0, 1]$. For convenience, let $Z \subseteq H$ denote the set of terminal histories.*

For convenience, let $H^i \subseteq H, i \in [N] \cup \{\mathcal{C}\}$ denote the set of histories with $P(h) = i$ and $\alpha(h)$ denote the set of valid actions at h, i.e., $\forall a \in \alpha(h)$, we have $(h, a) \in H$. Let $A = \max_h |\alpha(h)|$. A strategy σ^i for player i is a mapping from H^i to the distribution over valid actions, that is, $\sigma^i(h, a)$ is the probability of taking action a at $h \in H^i$. And a strategy profile σ consists of the strategies of all players in $[N]$, i.e., $\sigma = \{\sigma_i\}_{i\in[N]}$. We will use σ^{-i} to refer to the strategies of all players except i. In extensive games with imperfect information, we further divides H^i into information sets (infoset). Let $\mathcal{I}^i$ denote the set of infosets for player i. Since player i cannot distinguish $h_1, h_2 \in I \in \mathcal{I}^i$, so $\sigma^i(h_1), \sigma^i(h_2)$ must be the same. With a little abuse of notations, we use $\sigma^i(I)$ to denote the strategy on infoset I. Let $c^*(h, a), h \in H^{\mathcal{C}}$ denote the probability of $\mathcal{C}$ to take action a. For the convenience of notation, we use d to denote the corresponding (c, r). Let $u^i(h|\sigma, d^*)$ denote the expected reward of player $i \in [N]$ at history h under σ. For convenience, let $u^i(\sigma, d^*) = u^i(h_r|\sigma, d^*)$ where h_r is the root of H and $u^i(h|r^*)$ is the expected reward for $h \in Z$.

We will use $\pi_\sigma(h|d^*)$ to denote the probability of reaching h with σ and (c^*, r^*). It is easy to see that we can decompose $\pi_\sigma(h|d^*)$ into the product of the contribution of each player, that is, $\pi_\sigma(h|d^*) = \prod_{i\in[N]\cup\mathcal{C}} \pi^i_\sigma(h|d^*)$. And we will use $D(h)$ to refer to the depth of h in the game tree and $D^i(h)$ to refer to the number of h's ancestors whose player is i. Obviously, $D(h) = 1 + \sum_{i\in\{N\}\cup\{\mathcal{C}\}} D^i(h)$. And let $D = \max_h D(h)$ and $D^i = \max_h D^i(h)$.

Specifically, for a two-player zero-sum extensive game with imperfect information (TZIEG), $N = 2$ and $u^1(h) + u^2(h) = 0$ for all histories $h \in Z$.

Nash Equilibrium and exploitability: In a multi-agent system (MAS), a solution is often referred to a *Nash Equilibrium* (NE) (Osborne & Rubinstein, 1994). In a TZIEG, $\sigma = (\sigma^1, \sigma^2)$ is a NE if and only if $u^i(\sigma|d^*) = \max_{\sigma^{*,i}}(\sigma^{*,i}, \sigma^{-i}|d^*)$. In this work, we focus on approximating NE. More specifically, in TZIEGs, the approximation error of $\sigma = (\sigma^1, \sigma^2)$ is usually measured by its exploitability:

$$expl(\sigma|d^*) = \max_{\sigma^{*,1}} u^1(\sigma^{*,1}, \sigma^2|d^*) + \max_{\sigma^{*,2}} u^2(\sigma^1, \sigma^{*,2}|d^*) \tag{1}$$

In MARL, d^* is not known to the players, so players have to interact with the environment to gain knowledge about d^*. We assume players 1 and 2 can repeatedly play the game and collect observations in order to compute an approximate NE. Specifically, the collected data include the path $\mathcal{P}$ from the root of H to the terminal history, say h_z, and the reward. Thus, we observe samples from $c^*(h)$ for all $h \in \mathcal{P}, P(h) = \mathcal{C}$ and the sample from $r^{*,i}(h_z)$. We summarize the process in Alg. 1.

Algorithm 1 MARL for TZIEGs

Input: A TZIEG with unknown c^* and r^*.
while Not End **do**
  Compute an interaction strategy σ.
  Simulate σ in the environment to get observations about c^* and r^*.
end while
Output an approximate Nash Equilibrium $\bar{\sigma}$.

2.2 Posterior sampling for reinforcement learning (PSRL)

PSRL provides a framework under the Bayesian setting, where the environment is drawn from a given prior distribution. The process of PSRL can be decomposed into two steps: (1) estimating the parameters of the underlying environment with a posterior distribution; (2) sampling one environment from the posterior and computing strategies for agents according to the sampled environment. The computed strategies are used to interact with the underlying environment to collect data, so we call them interaction strategies. The two steps are repeated.

We also consider the TZIEG setting where the chance player and the reward functions follow a prior distribution $\mathbb{P}_0$. That is, the underlying c^* and r^* (i.e. d^*) are sampled from $\mathbb{P}_0(c, r)$. After playing t games, players collect some samples from c^* and r^* and they can get the posterior distribution, denoted as $\mathbb{P}_t$. For example, in the case where $r^*(h)$ is a Bernoulli distribution and its prior is a Beta distribution, the posterior distribution $\mathbb{P}_t(r)$ is also a Beta distribution. Similarly if the prior for c^* is a Dirichlet distribution, then $\mathbb{P}_t(c)$ is a Dirichlet distribution.

2.3 Counterfactual regret minimization (CFR)

Counterfactual regret minimization (CFR) (Zinkevich et al., 2008) is the state-of-the-art algorithm to solve TZIEGs when d^* is known. CFR is a self-play algorithm, which generates a sequence of strategy profiles, $\{\sigma_t\}_{t=1}^T$, by minimizing the following regrets:

$$R_T^{*,i} = \max_{\sigma^i} \sum_{t=1}^{T} u^i(\sigma^i, \sigma_t^{-i}|d^*) - \sum_{t=1}^{T} u^i(\sigma_t|d^*)$$

For convenience, we write $\bar{\sigma}_T = \frac{1}{T}\sum_{t=1}^T \sigma_t$ if $\bar{\sigma}_T^i(I) = \frac{\sum_{i,t} \pi_{\sigma_t}^i(I)\sigma_t^i(I)}{\sum_{i,t} \pi_{\sigma_t}^i(I)}$. One important observation is that (Zinkevich et al., 2008), in a TZIEG:

$$expl(\bar{\sigma}_T|d^*) = \frac{1}{T}(R_T^{*,1} + R_T^{*,2}) \tag{2}$$

Thus, minimizing R_T^* leads to the NE. CFR is an important sub-procedure in our algorithm.

Our algorithm is built in the frameworks of CFR and PSRL. In each round, we sample a d_t from $\mathbb{P}_t$ and then apply CFR to d_t. The main challenge on combining CFR and PSRL is how to interact with the environment. When there is only one player, PSRL selects the optimal strategy with respect to the sampled environment. But when there are multi-players, what is the optimal strategy is not clear. Later, we will show that we can use a time complexity only linear to the size of the game tree to compute an interaction strategy which leads to a convergence rate of $O(\sqrt{\log(T)/T})$.

3 METHOD

Algorithm 2 CFR-PSRL

while $t < T$ **do**
  Sample a chance player d_t from the posterior $\mathbb{P}_t$
  for all $i \in \{1, 2\}, I \in \mathcal{D}^i$ **do**
    Sample $d_t \sim \mathbb{P}_t$.
    Select σ_t by exploiting CFR to minimize the regret: $\max_{\sigma^i} \sum_{t \leq T} u^i(\sigma^i, \sigma_t^{-i}|d_t) - \sum_{t \leq T} u^i(\sigma_t|d_t)$.
  end for
  Select a sequence of interaction strategies to simulate to gather data and compute $\mathbb{P}_t$.
end while
Output: $\bar{\sigma} = \frac{1}{t} \sum_{t=1}^{T} \sigma_t$.

In this section, we introduce our method. We adopt the framework in Alg. 1 to develop our algorithm by combining PSRL and CFR, as well as designing the interaction strategies and procedure of computing the approximate NE.

Our algorithm is presented in Alg. 2. To compute the approximate NE, we adopt a CFR algorithm to minimize the following regret:

$$\hat{R}_T^i = \max_{\sigma^i} \sum_{t \leq T} u^i(\sigma^i, \sigma_t^{-i}|d_t) - \sum_{t \leq T} u^i(\sigma_t|d_t). \tag{3}$$

where d_t is sampled from $\mathbb{P}_t$. And then we output $\bar{\sigma} = \frac{1}{T} \sum_{t \leq T} \sigma_t$. Obviously, simply minimizing $\hat{R}_T$ will not make $expl(\bar{\sigma}|d^*)$ small, as d^* can be very different with d_t, so we need the interaction strategy to be efficient enough to make sure the difference between d_t and d^* is relatively small. The following equation establishes a relation between $expl(\bar{\sigma}|d^*)$ and $\hat{R}_T^i$:

$$expl(\bar{\sigma}|d^*) = \frac{1}{T}(\hat{R}_T^1 + \hat{R}_T^2 + \sum_{i \in \{1,2\}} \sum_{t \leq T} (u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T'^i, \sigma_t^{-i}|d_t))). \tag{4}$$

where $\sigma_T^{*,i} = \arg\max_{\sigma^i} \sum_{t \leq T} u^i(\sigma_T^i, \sigma_t^{-i}|d^*)$ and $\sigma_T'^i = \arg\max_{\sigma^i} \sum_{t \leq T} u^i(\sigma_T^i, \sigma_t^{-i}|d_t)$. For convenience, let $\mathcal{G}_T^i = \frac{1}{T} \sum_{t \leq T} (u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T'^i, \sigma_t^{-i}|d_t)))$.

The remaining challenge is to design interaction strategies to minimize $\mathcal{G}^i$. In round t, we first draw $\tilde{c}_t, \tilde{r}_t \sim \mathbb{P}_t$. Also, we denote $\tilde{d}_t = (\tilde{c}_t, \tilde{r}_t)$. And then for $i \in \{1, 2\}$, we compute

$$\tilde{\sigma}_t^i = \arg\max_{\sigma^i} \sum_{t'=1}^{t} \left(u^i(\sigma^i, \sigma_{t'}^{-i}|\tilde{d}_t) - u^i(\sigma^i, \sigma_{t'}^{-i}|d_{t'}) \right). \tag{5}$$

Interaction strategy: We adopt the following interaction strategies:

$$\hat{\sigma}_{1,T} = (\tilde{\sigma}_T^1, \sigma_T^2) \qquad \text{and} \qquad \hat{\sigma}_{2,T} = (\tilde{\sigma}_T^2, \sigma_T^1) \tag{6}$$

The computation of $\tilde{\sigma}$ can be implemented in time $O(|H|)$. With the interaction strategies $(\tilde{\sigma}_T^1, \sigma_T^2)$ and $(\tilde{\sigma}_T^2, \sigma_T^1)$, we can prove the following bound on $expl(\bar{\sigma})$.

Theorem 1. *Let* $\xi^i = \sum_{j=1}^{D} \sqrt{\max_{\sigma^i} \sum_{I \in \mathcal{I}^i, D(I)=j} \pi_{\sigma^i}^i(I)}$ *denote a game-dependent parameter. If the true game is sampled from a prior* $\mathbb{P}_0$ *over the chance player nodes and terminal nodes, then*

for $\bar{\sigma}_T$ computed by Alg. 2, we have

$$\begin{aligned}\frac{1}{T}(\hat{R}_T^1 + \hat{R}_T^2) =& O\left(\frac{1}{T}\left((\xi^1+\xi^2)\sqrt{AT}\right)\right),\\ \mathcal{G}_T^i =& O\left(\frac{1}{T}\left(\sqrt{|Z|T\ln(|Z|T)} + \sqrt{|H^{\mathcal{C}}|D^{\mathcal{C}}AT\ln(|H^{\mathcal{C}}|T)}\right)\right),\\ \mathbb{E}_{d^*} expl(\bar{\sigma}_T|d^*) =& O\left(\frac{1}{T}\left((\xi^1+\xi^2)\sqrt{AT} + \sqrt{|Z|T\ln(|Z|T)} + \sqrt{|H^{\mathcal{C}}|D^{\mathcal{C}}AT\ln(|H^{\mathcal{C}}|T)}\right)\right).\end{aligned}$$

The present theorem is significant at least in the following aspects.

Firstly, the per round running time is linear to the size of game tree and the bound is sublinear to T. Thus, we can expect our algorithm to reach a certain approximate error in a finite time.

Secondly, our theorem holds for any prior distribution over d^*. In practical TZIEGs, it is possible that the priors for h_1 and h_2, $h_1, h_1 \in H^{\mathcal{C}}$, are independent. Our theorem and algorithms can also be applied to such situations.

Lastly, our interaction strategies $\hat{\sigma}_{1,T}$ and $\hat{\sigma}_{2,T}$ only contribute to the bound for $\mathcal{G}_T^i$, which can be treated as the error for interaction strategy's exploring the environment. If we apply PSRL to a single-agent tree game, the Bayesian regret might be considered as some error caused by interacting with the environment. Using the analysis in (Osband et al., 2013), we can get that PSRL enjoys an averaged Bayesian regret bound of order $O(\sqrt{|Z|\ln(|Z|T)/T} + \sqrt{|H^{\mathcal{C}}|D^{\mathcal{C}}A\ln(|H^{\mathcal{C}}|T)/T})$ for a general prior. Therefore, our bound for $\mathcal{G}_T^i$ has a comparable order to the bound for the average Bayesian regret in PSRL.

3.1 Proof sketch of Theorem 1

Before diving into details, we introduce some additional notations. For episode t, we generate two trajectories by interacting with the environment. More specifically, we use $\mathcal{T}_{i,t}$ ($i \in \{1,2\}$) to denote the trajectory generated by $\hat{\sigma}_{i,t}$ in environment d^*. We use $\mathbb{E}_{\mathcal{T}_{i,t}}$ to denote the expectation over all trajectories for episode t. Then we denote $\mathcal{T}_{i,t}^{\mathcal{C}} = \{h_{1,t}^{\mathcal{C}}, h_{2,t}^{\mathcal{C}}, ..., h_{m_{i,t},t}^{\mathcal{C}}\}$ the trajectory for the chance player in episode t, and here $m_{i,t}$ denotes the length of $\mathcal{T}_{i,t}^{\mathcal{C}}$. Furthermore, we denote the terminal node for episode t as $z_{i,t}$. Besides, we denote the collection of $\mathcal{T}_{1,1}, \mathcal{T}_{2,1}..., \mathcal{T}_{1,t-1}, \mathcal{T}_{2,t-1}$ and the related rewards as $\mathcal{H}_t$, which represents all the observations before episode t. For each history h, we further use $n_t(h)$ to denote the count that h has been visited in $\mathcal{H}_t$.

Below we give the key part for the proof. Obviously, we need to bound the regret of CFR, i.e., $\hat{R}_T^i$, and $\mathcal{G}_T^i$. We can directly apply the technique in (Neil, 2018) to bound $\hat{R}_T^i$. Now we show the key part for bounding $\mathcal{G}_T^i$.

With straight-forward calculations, we have:

$$\begin{aligned}\mathcal{G}_T^i \leq& \frac{1}{T}\sum_{t\leq T}(u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d_t)))\\ \leq& \frac{1}{T}\max_{\sigma^i}\sum_{t\leq T}(u^i(\sigma^i, \sigma_t^{-i}|d^*) - u^i(\sigma^i, \sigma_t^{-i}|d_t))).\end{aligned}$$

And then, in Lemma 1 and 2, we decompose the bound into weighted sum of $|c^*(h) - c_t(h)|$ and $|r^*(h) - r_t(h)|$. And soon later we will show how to minimize each term by interaction.

Lemma 1. *With $\hat{\sigma}$ defined in Eq.* (6)*, we have:*

$$\begin{aligned}\mathbb{E}_{\mathcal{H}_T}\left\{\mathbb{E}_{d^*}\left[\mathcal{G}_T^i\middle|\mathcal{H}_T\right]\right\} \leq& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i, \sigma_t^{-i}|\tilde{d}_t) - u^i(\tilde{\sigma}_t^i, \sigma_t^{-i}|d^*)\right]\middle|\mathcal{H}_t\right\}\\ &+ \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i, \sigma_t^{-i}|d^*) - u^i(\tilde{\sigma}_t^i, \sigma_t^{-i}|d_t)\right]\middle|\mathcal{H}_t\right\}. \quad (7)\end{aligned}$$

Lemma 1 decomposes the expectation of $\mathcal{G}_T^i$ into two terms, representing the difference between d^* and $\tilde{d}_t$ and the difference between d^* and d_t. Below we give an intuitive sketch for bounding the first term $u^i(\tilde{\sigma}_t^i, \sigma_t^{-i}|\tilde{d}_t) - u^i(\tilde{\sigma}_t^i, \sigma_t^{-i}|d^*)$.

Lemma 2. *With $\mathbb{E}_{\mathcal{T}_{i,t}}$ denoting the expectation over trajectories, the following inequality holds*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|\tilde{d}_t)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d^*)\right]\Big|\mathcal{H}_t\right\}\\
\leq&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{j=1}^{m_{i,t}}\sum_{a\in\alpha(h)}|\tilde{c}_t(h_{j,t}^{\mathcal{C}},a)-c^*(h_{j,t}^{\mathcal{C}},a)|\right]\Big|\mathcal{H}_t\right\}\\
&+\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[u^i(s_t|\tilde{r}_t)-u^i(z_{i,t}|r^*)\right]\Big|\mathcal{H}_t\right\}. \qquad (8)
\end{aligned}$$

According to the definition of the expectation $\mathbb{E}_{\tilde{d},d^*}\mathbb{E}_{\mathcal{T}_{i,t}}$, we can see that Eq. (8) is a weighted sum of $|c_t(h) - c^*(h)|$ and $|u^i(h|\tilde{r}_t) - u^i(h|r^*)|$. Recall that $u^i(h|r)$ refers to the expectation of $r(h)$ for player i. Intuitively, we can use concentration bound on $|c_t(h) - c^*(h)|$, so that for h with a large weight, we should visit it for more times. Notice that the weight in Eq. (8) is essentially the probability of reaching h under our interaction strategy $\hat{\sigma}$ and the real environment c^*. Hence if we use $\hat{\sigma}$ to interact with the environment, we can expect our algorithm can visit h with large weight for sufficient times.

To simplify the derivation, we tentatively assume that $\tilde{d}_t$ and d^* are identically distributed for nodes $h_{j,i}^{\mathcal{C}}$ and $z_{i,t}$ conditioning on $\mathcal{H}_t$. That is, for any node h, with Pr referring to the probability of some event, we here assume that

$$Pr(d^*|\mathcal{H}_t, h) = Pr(\tilde{d}_t|\mathcal{H}_t, h).$$

In fact this assumption fails when h is reached, because the probability to reach h is influenced by d^* and $\tilde{d}$. We will remove this assumption and provide a rigorous proof in Appendix A. For $(h_{j,i}^{\mathcal{C}}, a)$ and $z_{i,t}$, we can insert the empirical mean estimations $\bar{c}_t(h_{j,i}^{\mathcal{C}}, a)$ and $\bar{u}_t^i(z_{i,t})$ and use the frequentists' concentration bound (Hoeffding, 1994; Weissman et al., 2003). Then for any $\delta \in (0, 1)$, we have the following inequalities:

$$\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{a\in\alpha(h_{j,t}^{\mathcal{C}})}|\tilde{c}(h_{j,t}^{\mathcal{C}},a)-c^*(h_{j,t}^{\mathcal{C}},a)|\right]\Big|\mathcal{H}_t\right\}\leq\mathbb{E}_{\mathcal{H}_t}\left[2\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h_{j,t}^{\mathcal{C}}),1)}}\Big|\mathcal{H}_t\right]+2|H^{\mathcal{C}}|\delta,$$

$$\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[u^i(h|\tilde{r}_t)-u^i(h|r^*)\right]\Big|\mathcal{H}_t\right\}\leq\mathbb{E}_{\mathcal{H}_t}\left[2\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\Big|\mathcal{H}_t\right]+4|Z|\delta.$$

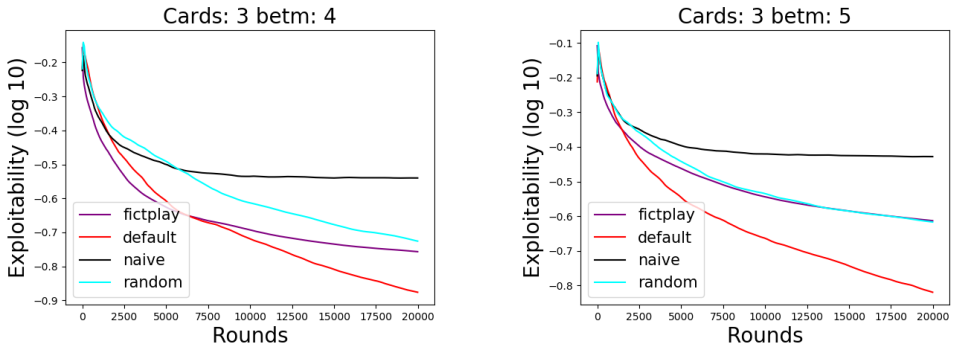
Then for a history $h \in Z \cup H^{\mathcal{C}}$, we have $\sum_{n=i}^{n_t(h)}\sqrt{1/i} \leq \sqrt{n_t(h)}$. Then we use the Jensen's inequality to the summation over Z and $H^{\mathcal{C}}$ to get the below bound:

$$\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|\tilde{d}_t)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d^*)\right]\Big|\mathcal{H}_t\right\}=O(\sqrt{|Z|T\ln(|Z|T)}+\sqrt{|H^{\mathcal{C}}|D^{\mathcal{C}}AT\ln(|H^{\mathcal{C}}|T)}).$$

We can apply the same method to $\sum_{t=1}^{T}\mathbb{E}_{H_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d^*)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d_t)\right]\Big|\mathcal{H}_t\right\}$ and finish the proof of the theorem 1.

4 Related Work

Fictitious Play: Fictitious play (FP) (Brown, 1951) is another popular algorithm for approximating NE in two-player zero-sum games. In FP, the agent takes the best response to the average strategy



(a) Leduc-4 (b) Leduc-5

Figure 1: Results for different algorithms on variants of Leduc-4 and Leduc-5.

of its opponent. Heinrich et al. (2015) extends FP to TZIEGs. Though it maybe easier to combine FP with other machine learning techniques than CFR, when the chance player is known, the convergence rate of FP is usually worse than CFR variants.

MDP: SARL problems is often formalized as the Markov Decision Process (MDP). In the simplest MDP with no transitions, i.e. the Multi-armed bandit problems, the problem-dependent regret upper bound of PSRL (also named Thompson Sampling in bandit problems) is carefully analyzed (Agrawal & Goyal, 2017). The problem-dependent bounds for general MDP is still an open problem. Besides PSRL, there is another kind of provable algorithms for MDP (Jaksch et al., 2010; Azar et al., 2013) following the Optimal in the Face of Uncertainty principle. They estimate the uncertainty of the underlying MDP and then use the currently optimal policy to interact with the environment.

Stochastic Games: the stochastic game (Littman, 1994) is also one kind MAS. In a stochastic game, players take actions at each state and then the environment transits to a new state and returns immediate rewards. Nash Q-learning (Hu & Wellman, 2003) converges to approximate NE by extending Q-learning to games, but it lacks finite-time analysis. Some other work (Szepesvári & Littman, 1996; Perolat et al., 2015; Wei et al., 2017) concentrates on two-player zero-sum stochastic games in RL setting. This kind of games don't involve imperfect information, and this makes them different from TZIEG.

5 Experiments

To empirically evaluate our algorithm, we test it on imperfect-information poker games. In this section, we first introduce our baseline methods and then present the details of the games. Finally, we show the results.

We choose two kinds of methods as our baseline methods. The first one is Fictitious Play (FP), an algorithm for solving MARL. Thus we can compare the performance of our algorithm and existing FP. We choose two variants of our algorithm as the other kind of baselines, which is used to compare different choices of interaction strategies. Details of baselines are given below:

- Fictitious Play (FP): FP is another popular algorithm to solve games in the RL setting. In FP, when d^* is known, each player chooses the best response of its opponent's average strategy. When d^* is not known, we need other RL algorithms to learn the best response. Heinrich et al. (2015) suggests to use a Fitted-Q iteration (FQI) algorithm (Ernst et al., 2005) to learn the best response. However, to use FQI, we have to specify the learning step size and the size of replay memory in advance, which is not needed in our algorithm. Instead, we use a combination of FP and PSRL: In round t, we compute player i's best response under $d_t \sim \mathbb{P}_t$, that is, $\arg\max_{\sigma^i} \sum_{t'<t} u^i(\sigma^i, \sigma_{t'}^{-i}|d_t)$.

- Variants of Alg. 2: Though we proved the convergence of Alg. 2 with interaction strategy $(\tilde{\sigma}^i, \sigma^{-i})$, the proposed method does not necessarily work well in practice. In our experiments, we evaluate three interaction strategies: 1) Random: the players take actions randomly; 2) Naive: the players use the output of the CFR procedure, i.e., σ_t, to interact with the environment; 3) the interaction strategies in Eq. (6).

We test these algorithms on variants of Leduc Hold'em poker (Southey et al., 2012) which is widely used in imperfect-information game solving. We generate games by keeping the tree structure of the Leduc Hold'em poker and replacing c and r by randomly generated functions. More specifically, when generating the tree structure, to control the sizes of the generated game tree, we restrict each player not to bid more than 4 or 5 times the big blind. The numbers of histories in the generated games are 9435 and 34776 respectively. The reward function $r^i(h)$ is a binary distribution. With a probability p the value of $r(h)$ is -1 and with probability $1-p$, the value is 1. The prior $\mathbb{P}_0(r(h))$ is a uniform $[0,1]$ distribution over parameter p. Let e^d denote the vector in $\mathbb{R}^d$ with every element is 1. $c(h)$ is sampled from $Dirichlet(e^{|A(h)|})$.

We generate 20 variants for Leduc(4) and Leduc(5) respectively. And on each generated game, each algorithm updates its strategies for 10000 times, and after each update, it interacts with the environment for 2 rounds. The result is in Fig. 1. As the figure shows, the exploitability of naive CFR fails to decrease after 10000 rounds on both Leduc 4 and 5. This might be caused by the lack of efficient exploration of the environment. Random interaction and FP can gradually decrease the exploitability, but our algorithm decrease at a higher speed. Thus the empirical result shows that our algorithm outperforms baselines on the two games.

6 Conclusion and discussion

In this work, we consider the problem of posterior sampling for TZIEGs, which is a class of multi-agent reinforcement learning problems. By a novel design of interaction staregies, we combine PSRL and CFR and present a provably convergent algorithm for TZIEGs. Our algorithm empirically works well. There is a large room to improve the result in the future, at least from the following directions:

At first, our bound is a Bayesian bound describing the expected performance. Considering one sample from the prior, Frequentists' methods such as UCBVI (Azar et al., 2013) also give a high probability regret bound for SARL of a similar order to PSRL. Further, comparing with the worst-case bound, the problem-dependent performance is much more important. Though it is possible that our method has a better performance on a specific TZIEG than the bound in Theorem 1, our algorithm is very possibly not the best in the sense of problem-dependent performance.

Secondly, our method heavily relies on the structure of TZIEGs and the solution concept Nash Equilibrium. Thus, further work is needed to extend posterior sampling to more complicated multi-agent systems, such as stochastic games (Littman, 1994) and extensive games with more than two players.

References

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):30, 2017.

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

George W Brown. Iterative solution of games by fictitious play, 1951. *Activity Analysis of Production and Allocation (TC Koopmans, Ed.)*, pp. 374–376, 1951.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257, 2011.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *International Conference on Machine Learning*, pp. 805–813, 2015.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

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

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

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Burch Neil. Time and space: Why imperfect information games are hard. 2018.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? *arXiv preprint arXiv:1607.00215*, 2016.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning (ICML 2015)*, 2015.

Finnegan Southey, Michael P Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker. *arXiv preprint arXiv:1207.1411*, 2012.

Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pp. 943–950, 2000.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Csaba Szepesvári and Michael L Littman. Generalized markov decision processes: Dynamic-programming and reinforcement-learning algorithms. In *Proceedings of International Conference of Machine Learning*, volume 96, 1996.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pp. 4987–4997, 2017.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, pp. 1729–1736, 2008.

A PROOF FOR THEOREM1

Let let $\bar{\sigma} = \frac{1}{T}\sum_{t\leq T}\sigma_t$. We decompose the exploitability at episode T into the CFR regret and an extra exploration term:

$$expl(\bar{\sigma}|d^*) = \frac{1}{T}(\hat{R}_T^1 + \hat{R}_T^2 + \sum_{i\in\{1,2\}}\sum_{t\leq T}(u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T'^i, \sigma_t^{-i}|d_t))),$$

where $\hat{R}_T^i$ and $\sigma_T'^i$ are formally defined as:

$$\hat{R}_T^i = \max_{\sigma^i}\sum_{t\leq T}u^i(\sigma^i, \sigma_t^{-i}|d_t) - \sum_{t\leq T}u^i(\sigma_t|d_t),$$
$$\sigma_T'^i = \arg\max_{\sigma^i}\sum_{t=1}^{T}u^i(\sigma^i, \sigma_t^{-i}|d_t).$$

Here d_t are sampled from the posterior distribution $\mathbb{P}_t$. The proof for this decomposition is given below:

Proof. With straight-forward computations, we have:

$$\sum_{t\leq T}u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) = \sum_{t\leq T}(u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T'^i, \sigma_t^{-i}|d_t) + u^i(\sigma_T'^i, \sigma_t^{-i}|d_t))$$
$$= \sum_{i,t}(u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T'^i, \sigma_t^{-i}|d_t)) + \sum_{t\leq T}u^i(\sigma_t|d_t) + \hat{R}_T^i$$

Moreover, with $u^1(\sigma_t|d_t) + u^2(\sigma_t|d_t) = 0$ and the definition of $expl$, we finish the proof.

□

By directly applying the result in (Neil, 2018), we can upper bound the CFR regret with

$$\hat{R}_T^i \leq \frac{1}{T}\left(\xi^i\sqrt{AT}\right),$$

where $\xi^i = \sum_{j=1}^{D}\sqrt{|B^i(j)|}$.

For convenience, let $\mathcal{G}_T^i = \frac{1}{T}\sum_{t\leq T}(u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T'^i, \sigma_t^{-i}|d_t)))$. We can use the standard analysis for CFR (Zinkevich et al., 2008; Neil, 2018) to bound $\hat{R}_T^i$. Thus, we only need to bound $\mathcal{G}^i$.

Obviously, $\mathcal{G}_T^i$ depends on the difference between c^* and c_t. So that we need to design a suitable interaction strategy to make sure that $\mathcal{G}_T^i$ is small. We upper bound $\mathcal{G}_T^i$ with

$$\mathcal{G}_T^i \leq \frac{1}{T}\sum_{t\leq T}(u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d^*) - u^i(\sigma_T^{*,i}, \sigma_t^{-i}|d_t)))$$
$$\leq \frac{1}{T}\max_{\sigma^i}\sum_{t\leq T}(u^i(\sigma^i, \sigma_t^{-i}|d^*) - u^i(\sigma^i, \sigma_t^{-i}|d_t))).$$

We select the interaction strategy as follows: draw $\tilde{d}_t \sim \mathbb{P}_t$. For $i \in \{1, 2\}$, compute

$$\tilde{\sigma}_t^i = \arg\max_{\sigma^i}\sum_{t'\leq t}u^i(\sigma^i, \sigma_t^{-i}|\tilde{d}_t) - u^i(\sigma^i, \sigma_t^{-i}|d_t) \tag{9}$$

And then we use $(\tilde{\sigma}_t^1, \sigma_t^2)$ and $(\tilde{\sigma}_t^2, \sigma_t^1)$ to interact with the environment. Following lemma provides an upper bound on $\mathcal{G}_t^i$. Then we get below lemma:

Lemma 3. *With $\tilde{\sigma}_t^i$ defined in Eq.* (9)*, we have:*

$$\mathbb{E}_{H_T}\left\{\mathbb{E}_{d^*}\left[\mathcal{G}_T^i\middle|H_T\right]\right\} \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{H_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|\tilde{d}_t)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d^*)\right]\middle|\mathcal{H}_t\right\} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{H_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d^*)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d_t)\right]\middle|\mathcal{H}_t\right\}. \quad (10)$$

This lemma decompose $\mathcal{G}_T^i$ into two terms, which can be bounded with careful analysis of the posterior distribution. We first give the proof for this lemma.

Proof. We have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_T}\left\{\mathbb{E}_{d^*}\left[\max_{\hat{\sigma}^i}\left(\sum_{t=1}^{T}(u^i(\hat{\sigma}^i,\sigma_t^{-i}|d^*)-u^i(\hat{\sigma}^i,\sigma_t^{-i}|d_t))\right)\right]\middle|\mathcal{H}_T\right\}\\
=&\mathbb{E}_{H_T}\left\{\mathbb{E}_{\tilde{d}_T}\left[\max_{\tilde{\sigma}^i}\left(\sum_{t=1}^{T}(u^i(\tilde{\sigma}^i,\sigma_t^{-i}|\tilde{d}_T)-u^i(\tilde{\sigma}^i,\sigma_t^{-i}|d_t))\right)\right]\middle|\mathcal{H}_T\right\}\\
=&\mathbb{E}_{\mathcal{H}_T}\left\{\mathbb{E}_{\tilde{d}_T}\left[\sum_{t=1}^{T}(u^i(\tilde{\sigma}_T^i,\sigma_t^{-i}|\tilde{d}_T)-u^i(\tilde{\sigma}_T^i,\sigma_t^{-i}|d_t))\right]\middle|\mathcal{H}_T\right\}\\
=&\mathbb{E}_{\mathcal{H}_T}\left\{\mathbb{E}_{\tilde{d}_T}\left[\sum_{t=1}^{T-1}(u^i(\tilde{\sigma}_T^i,\sigma_t^{-i}|\tilde{d}_T)-u^i(\tilde{\sigma}_T^i,\sigma_t^{-i}|d_t))\right]\middle|\mathcal{H}_T\right\}\\
&+\mathbb{E}_{\mathcal{H}_T}\left\{\mathbb{E}_{\tilde{d}_T}\left[u^i(\tilde{\sigma}_T^i,\sigma_T^{-i}|\tilde{d}_T)-u^i(\tilde{\sigma}_T^i,\sigma_T^{-i}|d_T)\right]\middle|\mathcal{H}_T\right\}\\
\leq&\mathbb{E}_{\mathcal{H}_T}\left\{\mathbb{E}_{\tilde{d}_T}\left[\max_{\tilde{\sigma}^i}\left(\sum_{t=1}^{T-1}(u^i(\tilde{\sigma}^i,\sigma_t^{-i}|\tilde{d}_T)-u^i(\tilde{\sigma}^i,\sigma_t^{-i}|d_t))\right)\right]\middle|\mathcal{H}_T\right\}\\
&+\mathbb{E}_{H_T}\left\{\mathbb{E}_{\tilde{d}_T}\left[u^i(\tilde{\sigma}_T^i,\sigma_T^{-i}|\tilde{d}_T)-u^i(\tilde{\sigma}_T^i,\sigma_T^{-i}|d_T)\right]\middle|\mathcal{H}_T\right\}\\
=&\mathbb{E}_{\mathcal{H}_{T-1}}\left\{\mathbb{E}_{\tilde{d}_{T-1}}\left[\sum_{t=1}^{T-1}(u^i(\tilde{\sigma}_{T-1}^i,\sigma_t^{-i}|\tilde{d}_{T-1})-u^i(\tilde{\sigma}_{T-1}^i,\sigma_t^{-i}|d_t))\right]\middle|\mathcal{H}_{T-1}\right\}\\
&+\mathbb{E}_{\mathcal{H}_T}\left\{\mathbb{E}_{\tilde{d}_T}\left[u^i(\tilde{\sigma}_T^i,\sigma_T^{-i}|\tilde{d}_T)-u^i(\tilde{\sigma}_T^i,\sigma_t^{-i}|d_T)\right]\middle|\mathcal{H}_T\right\}\\
\leq&\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|\tilde{d}_t)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d_t)\right]\middle|\mathcal{H}_t\right\}\\
=&\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|\tilde{d}_t)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d^*)\right]\middle|\mathcal{H}_t\right\}\\
&+\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d^*)-u^i(\tilde{\sigma}_t^i,\sigma_t^{-i}|d_t)\right]\middle|\mathcal{H}_t\right\}.
\end{aligned}$$

The first equality holds since d^* and $\tilde{d}_T$ are identical distributed conditioning on $\mathcal{H}_t$.

Therefore, we finish the proof. □

Then we only to give upper bounds for the two terms. Also mentioned in sec.3.1, we introduce some additional notations. For episode t, we generate two trajectories by interacting with the environment. More specifically, we use $\mathcal{T}_{i,t}$ ($i \in \{1,2\}$) to denote the trajectory generated by $\hat{\sigma}_{i,t}$ with d^*. We use $\mathbb{E}_{\mathcal{T}_{i,t}}$ to denote the expectation over all trajectories for episode t. Then we denote $\mathcal{T}_{i,t}^{\mathcal{C}} = \{h_{1,t}^{\mathcal{C}}, h_{2,t}^{\mathcal{C}}, ..., h_{m_{i,t},t}^{\mathcal{C}}\}$ the trajectory for the chance player in episode t, and here $m_{i,t}$ denotes

the length of $\mathcal{T}^{\mathcal{C}}_{i,t}$. Furthermore, we denote the terminal node for episode t as $z_{i,t}$. Besides, we denote the collection of $\mathcal{T}_{1,1}, \mathcal{T}_{2,1}..., \mathcal{T}_{1,t-1}, \mathcal{T}_{2,t-1}$ as $\mathcal{H}_t$, which represents all the observations before episode t. For each history h, we further use $n_t(h)$ to denote the count that h has been visited in $\mathcal{H}_t$.

Then we concentrate on the first term $\sum_{t=1}^{T} \mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|\tilde{d}_t) - u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|d^*)\right]\Big|\mathcal{H}_t\right\}$ and the second term has similar proof. Since the strategy tuple is the same for the two utilities, we can decompose their difference with below lemma.

Lemma 4. *With $\mathbb{E}_{\mathcal{T}_{i,t}}$ denoting the expectation over trajectories, the following inequality holds*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|\tilde{d}_t) - u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|d^*)\right]\Big|\mathcal{H}_t\right\}\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{j=1}^{m_{i,t}}\sum_{a\in\alpha(h)}(\tilde{c}_t(h^{\mathcal{C}}_{j,t},a) - c^*(h^{\mathcal{C}}_{j,t},a))u^i(h^{\mathcal{C}}_{j,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t)\right]\Big|\mathcal{H}_t\right\}\\
&+\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[u^i(z_{i,t}|\tilde{r}_t) - u^i(z_{i,t}|r^*)\right]\Big|\mathcal{H}_t\right\}.
\end{aligned}$$

Proof. From the root node to $h^{\mathcal{C}}_{1,t}$, players take actions according to $(\tilde{\sigma}^i_t,\sigma^{-i}_t)$. Thus we should have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|\tilde{d}_t) - u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|d^*)\right]\Big|\mathcal{H}_t\right\}\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t) - u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,d^*)\right]\Big|\mathcal{H}_t\right\}\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{a\in\alpha(h^{\mathcal{C}}_{1,t})}(\tilde{c}_t(h^{\mathcal{C}}_{1,t},a)u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t) - c^*(h^{\mathcal{C}}_{1,t},a)u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,d^*))\right]\Big|\mathcal{H}_t\right\}\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{a\in\alpha(h^{\mathcal{C}}_{1,t})}(\tilde{c}_t(h^{\mathcal{C}}_{1,t},a) - c^*(h^{\mathcal{C}}_{1,t},a))u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t)\right]\Big|\mathcal{H}_t\right\}\\
&+\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{a\in\alpha(h^{\mathcal{C}}_{1,t})}c^*(h^{\mathcal{C}}_{1,t},a)(u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t) - u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,d^*))\right]\Big|\mathcal{H}_t\right\}\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{a\in\alpha(h^{\mathcal{C}}_{1,t})}(\tilde{c}_t(h^{\mathcal{C}}_{1,t},a) - c^*(h^{\mathcal{C}}_{1,t},a))u^i(h^{\mathcal{C}}_{1,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t)\right]\Big|\mathcal{H}_t\right\}\\
&+\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[u^i(h^{\mathcal{C}}_{2,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t) - u^i(h^{\mathcal{C}}_{2,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,d^*)\right]\Big|\mathcal{H}_t\right\}\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{j=1}^{m_{i,t}}\sum_{a\in\alpha(h^{\mathcal{C}}_{j,t})}(\tilde{c}_t(h^{\mathcal{C}}_{j,t},a) - c^*(h^{\mathcal{C}}_{j,t},a))u^i(h^{\mathcal{C}}_{j,t}|\tilde{\sigma}^i_t,\sigma^{-i}_t,\tilde{d}_t)\right]\Big|\mathcal{H}_t\right\}\\
&+\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[u^i(z_{i,t}|\tilde{r}_t) - u^i(z_{i,t}|r^*)\right]\Big|\mathcal{H}_t\right\}.
\end{aligned}$$

Therefore we finish the proof. □

We upper bound the term $u^i(z_{i,t}|\tilde{r}_t) - u^i(z_{i,t}|r^*)$ first. We can refer to the technique of previous work in PSRL. Recall that in episode t, players reaches terminal node $z_{i,t}$ with a visited count $n_t(z_{i,t})$. We denote that $\bar{u}^i_t(z_{i,t})$ as the empirical mean of $u^i(z_{i,t}|r^*)$. Simply we insert $\bar{u}^i_t(z_{i,t})$ to get

$$u^i(z_{i,t}|\tilde{r}_t) - u^i(z_{i,t}|r^*) \le |u^i(z_{i,t}|\tilde{r}_t) - \bar{u}^i_t(z_{i,t})| + |\bar{u}^i_t(z_{i,t}) - u^i(z_{i,t}|r^*)|.$$

First we consider the second one $|\bar{u}^i_t(z_{i,t}) - u^i(z_{i,t}|r^*)|$. Conditioning on $r^*(z_{i,t})$, we can apply the Chernoff-Hoeffding bound (Hoeffding, 1994). For $\delta \in (0,1)$

$$Pr\left(|\bar{u}^i_t(z_{i,t}) - u^i(z_{i,t}|r^*)| \ge \sqrt{\frac{\ln(2/\delta)}{2\max(n_t(z_{i,t}),1)}}\Big|r^*(z_{i,t})\right) \le \delta, \tag{11}$$

where Pr denote the probability.

Then we use the above inequality to get below lemma:

Lemma 5.

$$\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[|\bar{u}_t^i(z_{i,t})-u^i(z_{i,t}|r^*)|\right]\Big|\mathcal{H}_t\right\}\leq\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}+2|Z|\delta.$$

Proof. Notice that Eq. 11 holds conditioning on $r^*(z_{i,t})$ and the expectation is taken over the prior $\mathbb{P}_0$. Then we need to carefully apply the Eq. 11. For the convenience of notation, we use $\pi_t(h|d^*)$ to represent $\pi_{\tilde{\sigma}_t^i,\sigma_t^{-i}}(h|d^*)$. We further use $\mathbb{I}(\cdot)$ to indicate the identical function. Then we expand the expectation into integration

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[|\bar{u}_t^i(z_{i,t})-u^i(z_{i,t}|r^*)|\right]\Big|\mathcal{H}_t\right\}\\
=&\sum_{z\in Z}\int|\bar{u}_t^i(z)-u^i(z|r^*)|\pi_t(z|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\
\leq&\sum_{z\in Z}\int\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(z),1)}}\pi_t(z|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\
&+\sum_{z\in Z}\int 2\mathbb{I}\left(|\bar{u}_t^i(z)-u^i(z|r^*)|\geq\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(z),1)}}\right)Pr(d^*|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,\mathcal{H}_t)\qquad(12)\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}\\
&+\sum_{z\in Z}\int 2\mathbb{I}\left(|\bar{u}_t^i(z)-u^i(z|r^*)|\geq\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(z),1)}}\right)Pr(\mathcal{H}_t|d^*)\mathbb{P}_0(d^*)d(d^*,\mathcal{H}_t)\\
\leq&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}+2|Z|\delta.
\end{aligned}$$

Therefore we finish the proof. □

For another term $|u^i(z_{i,t}|\tilde{r}_t)-\bar{u}_t^i(z_{i,t})|$, we can still apply Lemma 5 to get below lemma:

Lemma 6.

$$\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[|u^i(z_{i,t}|\tilde{r}_t)-\bar{u}_t^i(z_{i,t})|\right]\Big|\mathcal{H}_t\right\}\leq\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}+2|Z|\delta.$$

Proof. We can directly prove that

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[|u^i(z_{i,t}|\tilde{r}_t)-\bar{u}_t^i(z_{i,t})|\right]\Big|\mathcal{H}_t\right\}\\
=&\sum_{s\in Z}\int|u^i(s|\tilde{r}_t)-\bar{u}_t^i(s)|\pi_t(s|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\
\leq&\sum_{s\in Z}\int\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(s),1)}}\pi_t(s|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\
&+\sum_{s\in Z}\int 2\mathbb{I}\left(|u^i(s|\tilde{r}_t)-\bar{u}_t^i(s)|\geq\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(s),1)}}\right)Pr(\tilde{d}|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,\mathcal{H}_t) \qquad (13)\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}\\
&+\sum_{s\in Z}\int 2\mathbb{I}\left(|u^i(s|\tilde{r}_t)-\bar{u}_t^i(s)|\geq\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(s),1)}}\right)Pr(\mathcal{H}_t|d^*)\mathbb{P}_0(d^*)d(d^*,\mathcal{H}_t)\\
\leq&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}+2|Z|\delta.
\end{aligned}
$$

Since d^* and $\tilde{d}_t$ are identically distributed conditioning on $\mathcal{H}_t$, then we apply below equality to Eq. (13):

$$
\begin{aligned}
&\mathbb{I}\left(|u^i(s|\tilde{r}_t)-\bar{u}_t^i(s)|\geq\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(s),1)}}\right)Pr(\tilde{d}_t|\mathcal{H}_t)Pr(\mathcal{H}_t)\\
=&\mathbb{I}\left(|u^i(s|r^*)-\bar{u}_t^i(s)|\geq\sqrt{\frac{\ln(2/\delta)}{2\max(n_t(s),1)}}\right)Pr(d^*|\mathcal{H}_t)Pr(\mathcal{H}_t).
\end{aligned}
$$

Then we finish the proof. □

Hence we combine the results in Lemma 5 and 6 and get the conclusion that for any $\delta\in(0,1)$,

$$
\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[u^i(z_{i,t}|\tilde{r}_t)-u^i(z_{i,t}|r^*)\right]\Big|\mathcal{H}_t\right\}\leq\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[2\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}+4|Z|\delta.
$$

Using a pigeon-hole principle and choosing $\delta=1/(|Z|T)$, we have below lemma:

Lemma 7. *At episode T,*

$$
\mathbb{E}_{\mathbb{P}_0}\left[\sum_{t=1}^{T}u^i(z_{i,t}|\tilde{r}_t)-u^i(z_{i,t}|r^*)\right]=O(\sqrt{|Z|T\ln(|Z|T)}).
$$

Then we consider chance player node $h_{j,t}^{\mathcal{C}}$. We also denote $\bar{c}(h_{j,t}^{\mathcal{C}},a)$ as the empirical mean of chance player's probability to choose a at $h_{j,t}^{\mathcal{C}}$. Notice that the utility is bounded in $[-1,1]$. We have

$$
\begin{aligned}
&\sum_{a\in\alpha(h_{j,t}^{\mathcal{C}})}(\tilde{c}_t(h_{j,t}^{\mathcal{C}},a)-c^*(h_{j,t}^{\mathcal{C}},a))u^i(h_{j,t}^{\mathcal{C}}|\tilde{\sigma}_t^i,\sigma_t^{-i},\tilde{d}_t)\\
\leq&2\sum_{a\in\alpha(h_{j,t}^{\mathcal{C}})}|\tilde{c}_t(h_{j,t}^{\mathcal{C}},a)-c^*(h_{j,t}^{\mathcal{C}},a)|\\
\leq&2\sum_{a\in\alpha(h_{j,t}^{\mathcal{C}})}|\tilde{c}_t(h_{j,t}^{\mathcal{C}},a)-\bar{c}(h_{j,t}^{\mathcal{C}},a)|+2\sum_{a\in\alpha(h)}|\bar{c}(h_{j,t}^{\mathcal{C}},a)-c^*(h_{j,t}^{\mathcal{C}},a)|.
\end{aligned}
$$

Then conditioning on $c^*(h^{\mathcal{C}}_{j,t}, a)$, we use the concentration bound for L_1 norm (i.e. the deviation inequality (Weissman et al., 2003) to get that for $\delta \in (0,1)$

$$Pr\left(\sum_{a\in\alpha(h^{\mathcal{C}}_{j,t})} |\bar{c}(h^{\mathcal{C}}_{j,t},a) - c^*(h^{\mathcal{C}}_{j,t},a)| \geq \sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h^{\mathcal{C}}_{j,t}),1)}}\Big| c^*(h^{\mathcal{C}}_{j,t},a)\right) < \delta.$$

Similar to the analysis in r, we give below lemma:

Lemma 8.

$$\begin{aligned}&\sum_{a\in\alpha(h^{\mathcal{C}}_{j,t})} \mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[|\bar{c}(h^{\mathcal{C}}_{j,t},a) - c^*(h^{\mathcal{C}}_{j,t},a)|\right]\Big|\mathcal{H}_t\right\}\\ \leq&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h^{\mathcal{C}}_{j,t}),1)}}\right]\Big|\mathcal{H}_t\right\} + |H^{\mathcal{C}}|\delta.\end{aligned}$$

Proof. We use similar techniques to get

$$\begin{aligned}&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{a\in\alpha(h^{\mathcal{C}}_{j,t})}|\bar{c}(h^{\mathcal{C}}_{j,t},a) - c^*(h^{\mathcal{C}}_{j,t},a)|\right]\Big|\mathcal{H}_t\right\}\\ =&\sum_{h\in H^{\mathcal{C}}}\int\sum_{a\in\alpha(h)}|\bar{c}(h,a) - c^*(h,a)|\pi_t(h|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\ \leq&\sum_{h\in H^{\mathcal{C}}}\int\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\pi_t(s|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\ &+\sum_{h\in H^{\mathcal{C}}}\int\mathbb{I}\left(\sum_{a\in\alpha(h)}|\bar{c}(h,a) - c^*(h,a)| \geq \sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\right)Pr(d^*|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,\mathcal{H}_t)\end{aligned} \tag{14}$$

$$\begin{aligned}=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\right]\Big|\mathcal{H}_t\right\}\\ &+\sum_{h\in H^{\mathcal{C}}}\int\mathbb{I}\left(\sum_{a\in\alpha(h)}|\bar{c}(h,a) - c^*(h,a)| \geq \sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\right)Pr(\mathcal{H}_t|d^*)\mathbb{P}_0(d^*)d(d^*,\mathcal{H}_t)\\ \leq&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\} + |H^{\mathcal{C}}|\delta.\end{aligned}$$

Therefore we finish the proof. □

Once again, we get below lemma:

Lemma 9.

$$\begin{aligned}&\sum_{a\in\alpha(h^{\mathcal{C}}_{j,t})} \mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[|\tilde{c}(h^{\mathcal{C}}_{j,t},a) - \bar{c}(h^{\mathcal{C}}_{j,t},a)|\right]\Big|\mathcal{H}_t\right\}\\ \leq&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h^{\mathcal{C}}_{j,t}),1)}}\right]\Big|\mathcal{H}_t\right\} + |H^{\mathcal{C}}|\delta.\end{aligned}$$

Proof. We use similar techniques to get

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\mathbb{E}_{\mathcal{T}_{i,t}}\left[\sum_{a\in\alpha(h^{\mathcal{C}}_{j,t})}|\tilde{c}(h^{\mathcal{C}}_{j,t},a)-\bar{c}(h^{\mathcal{C}}_{j,t},a)|\right]\Big|\mathcal{H}_t\right\}\\
=&\sum_{h\in H^{\mathcal{C}}}\int\sum_{a\in\alpha(h)}|\tilde{c}(h,a)-\bar{c}(h,a)|\pi_t(h|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\
\leq&\sum_{h\in H^{\mathcal{C}}}\int\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\pi_t(s|d^*)Pr(d^*,d_t|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,d_t,\mathcal{H}_t)\\
&+\sum_{h\in H^{\mathcal{C}}}\int\mathbb{I}\left(\sum_{a\in\alpha(h)}|\tilde{c}(h,a)-\bar{c}(h,a)|\geq\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\right)Pr(\tilde{d}|\mathcal{H}_t)Pr(\mathcal{H}_t)d(d^*,\mathcal{H}_t)\\
=&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\right]\Big|\mathcal{H}_t\right\}\\
&+\sum_{h\in H^{\mathcal{C}}}\int\mathbb{I}\left(\sum_{a\in\alpha(h)}|\bar{c}(h,a)-c^*(h,a)|\geq\sqrt{\frac{2\ln(2^A/\delta)}{\max(n_t(h),1)}}\right)Pr(\mathcal{H}_t|d^*)\mathbb{P}_0(d^*)d(d^*,\mathcal{H}_t)\\
\leq&\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[\sqrt{\frac{2\log(2/\delta)}{\max(n_t(z_{i,t}),1)}}\right]\Big|\mathcal{H}_t\right\}+|H^{\mathcal{C}}|\delta.
\end{aligned}\tag{15}$$

Therefore we finish the proof. □

Next, we use a pigeon-hole principle and choosing $\delta=1/(|H^{\mathcal{C}}|T)$, we have below lemma:

Lemma 10. *At episode T,*

$$\mathbb{E}_{\mathbb{P}_0}\left[\sum_{t=1}^{T}\sum_{j=1}^{m_{i,t}}\sum_{a\in\alpha(h)}|\tilde{c}_t(h^{\mathcal{C}}_{j,t},a)-c^*(h^{\mathcal{C}}_{j,t},a)|\right]=O(\sqrt{|H^{\mathcal{C}}|D^{\mathcal{C}}AT\ln(|H^{\mathcal{C}}|T)}).$$

Therefore we use the conclusion in Lemma7 and 10 to get

$$\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|\tilde{d}_t)-u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|d^*)\right]\Big|\mathcal{H}_t\right\}=O(\sqrt{|Z|T\ln(|Z|T)}+\sqrt{|H^{\mathcal{C}}|D^{\mathcal{C}}AT\ln(|H^{\mathcal{C}}|T)}).$$

The similar proof can be applied to the second term to get the same upper bound by simply replacing $\tilde{d}_t$ with d_t:

$$\sum_{t=1}^{T}\mathbb{E}_{\mathcal{H}_t}\left\{\mathbb{E}_{\tilde{d}_t,d^*}\left[u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|d^*)-u^i(\tilde{\sigma}^i_t,\sigma^{-i}_t|d_t)\right]\Big|\mathcal{H}_t\right\}=O(\sqrt{|Z|T\ln(|Z|T)}+\sqrt{|H^{\mathcal{C}}|D^{\mathcal{C}}AT\ln(|H^{\mathcal{C}}|T)}).$$

Sum the analysis together, we get to the conclusion that

$$\mathbb{E}_{H_T}\left\{\mathbb{E}_{d^*}\left[\mathcal{G}^i_T\Big|H_T\right]\right\}=O(\sqrt{\frac{|Z|\ln(|Z|T)}{T}}+\sqrt{\frac{|H^{\mathcal{C}}|D^{\mathcal{C}}A\ln(|H^{\mathcal{C}}|T)}{T}})$$