Towards Provably Efficient Learning of Imperfect Information Extensive-Form Games with Linear Function Approximation

Canzhe Zhao [*1] **Shuze Chen**[1] **Weiming Liu**[2] **Haobo Fu**[2] **Qiang Fu**[2] **Shuai Li** [†1]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]Tencent AI Lab, Shenzhen, China

Abstract

Despite significant advances in learning imperfect information extensive-form games (IIEFGs), most existing theoretical guarantees are limited to IIEFGs in the tabular case. To permit efficient learning of large-scale IIEFGs, we take the first step in studying two-player zero-sum IIEFGs with linear function approximation. In particular, we consider linear IIEFGs in the formulation of partially observable Markov games (POMGs) with linearly parameterized rewards. To address the challenge that the underlying function approximation structure is difficult to directly apply due to the imperfect information of states, we construct the composite "feature vectors" for information set-action pairs. Based on this, we further propose a "least-squares loss estimator", which we call the *fictitious* least-squares loss estimator. Through integrating this estimator with the follow-the-regularized-leader (FTRL) framework, we propose the *fictitious* least-squares follow-the-regularized-leader ($\texttt{F}^2\texttt{TRL}$) algorithm, which achieves a provable $\widetilde{\mathcal{O}}(\lambda\sqrt{dH^2T})$ regret guarantee in the large T regime, where d is the ambient dimension of the feature mapping, H is the horizon length, λ is a "balance coefficient" and T is the number of episodes. At the core of the analysis of $\texttt{F}^2\texttt{TRL}$ is the leverage of our proposed new "balanced transition" over information set-action space. Additionally, we complement our results with an $\Omega(\sqrt{d\min(d,H)T})$ regret lower bound for this problem and conduct empirical evaluations across various environments, which corroborate the effectiveness of our algorithm.

1 INTRODUCTION

In imperfect information games (IIGs), players only have partial observations of the true state of the game. Particularly, the notion of imperfect-information extensive-form games (IIEFGs) [Kuhn, 1953] simultaneously enables imperfect information and the sequencing of players' moves, which thus characterizes a large amount of real-world imperfect information games including Poker [Heinrich et al., 2015, Moravčík et al., 2017, Brown and Sandholm, 2018], Bridge [Tian et al., 2020], Scotland Yard [Schmid et al., 2021] and Mahjong [Li et al., 2020, Kurita and Hoki, 2021, Fu et al., 2022]. There has been a voluminous amount of works on regret minimization or finding the Nash equilibrium (NE) [Nash Jr, 1950] in IIEFGs. When the full knowledge of the game is known, existing works solve IIEFGs by linear programming [Koller and Megiddo, 1992, Von Stengel, 1996, Koller et al., 1996], first-order optimization methods [Hoda et al., 2010, Kroer et al., 2015, 2018, Munos et al., 2020, Lee et al., 2021, Liu et al., 2022], and counterfactual regret minimization (CFR) [Zinkevich et al., 2007, Lanctot et al., 2009, Johanson et al., 2012, Tammelin, 2014, Schmid et al., 2019, Burch et al., 2019, Liu et al., 2022].

When the full knowledge of the game is not known a priori, the problem will be much more challenging and is typically tackled through *learning* from the random samples accrued during repeated playthroughs of the game. In this line of works, learning two-player zero-sum IIEFGs have been addressed using Monte-Carlo CFR methods [Lanctot et al., 2009, Farina et al., 2020, Farina and Sandholm, 2021] or equipping online mirror descent (OMD) and follow-the-regularized-leader (FTRL) frameworks with loss estimators [Farina et al., 2021, Kozuno et al., 2021, Bai et al., 2022, Fiegel et al., 2023]. Amongst these works, Bai et al. [2022] leverage OMD with "balanced exploration policies" to achieve the $\widetilde{\mathcal{O}}(\sqrt{H^3XAT})$ regret bound, where H is the horizon length, X is the cardinality of the information set space, A is the cardinality of the action space and T is the number of episodes. Notably, this regret upper bound

[*]The majority of this work was done during Canzhe Zhao's internship at Tencent AI Lab.

[†]Correspondence to: Shuai Li <shuaili8@sjtu.edu.cn>

Table 1: Comparisons of regret bounds with most related works studying IIEFGs with bandit feedback.

Algorithm	Setting	Regret
IXOMD [Kozuno et al., 2021]	Tabular IIEFGs	$\widetilde{\mathcal{O}}(HX\sqrt{AT})$
BalancedOMD/CFR [Bai et al., 2022]		$\widetilde{\mathcal{O}}(\sqrt{H^3XAT})$
BalancedFTRL [Fiegel et al., 2023]		$\widetilde{\mathcal{O}}(\sqrt{XAT})$
F^{2}TRL (this paper)	Linear IIEFGs	$\widetilde{\mathcal{O}}(\lambda H\sqrt{dT})$ [1]
Lower bound (this paper)		$\Omega(\sqrt{d\min(d,H)T})$

[1] An exponential term that approaches 1 for large enough T is omitted for simplicity. Please see Theorem 1 for details. The "balance coefficient" λ is formally defined in Section 4.

matches the information-theoretic lower bound on all parameters but H up to logarithmic factors. Subsequently, Fiegel et al. [2023] further improve the upper bound to $\widetilde{\mathcal{O}}(\sqrt{XAT})$, which has optimal dependence on all parameters up to logarithmic factors, using FTRL with "balanced transitions".

Though significant advances have emerged in learning two-player zero-sum IIEFGs, the existing regret bounds of all works have polynomial dependence on X and A. In practice, however, X or A might be prohibitively large, which makes these regret bounds and sample complexities vacuous. To cope with this issue, a common approach is *function approximation*, which approximates the observations on experienced information sets and actions with sharing parameters and generalizes experienced observations onto unseen information sets and actions. Indeed, for practitioners in the area of IIEFGs (*e.g.*, Moravčík et al. [2017], Brown et al. [2019]), function approximation using, for example, deep neural networks, has made significant progress in solving large-scale IIEFGs. Yet, the theoretical guarantees of learning IIEFGs with function approximation still remain open and we are still far from understanding them well. This naturally motivates us to ask the following question:

Does there exist a provably efficient algorithm for learning IIEFGs in the function approximation setting?

In this paper, we give an affirmative answer to the above question for IIEFGs with linear function approximation over rewards and known sequence-form transition probabilities. Specifically, we consider IIEFGs in the formulation of partially observable Markov games (POMGs) with linearly parameterized rewards in the bandit feedback setting, in which only the information sets instead of the underlying states of the game are observable. This problem is challenging in the sense that the feature corresponding to the current state is unknown since the current state itself is unknown and only imperfect information of the current state is revealed to the learner, which poses substantial difficulties in exploiting the linear structure of the reward functions. To address this problem so as to establish provably efficient algorithms for learning IIEFGs with linear function approximation, we make the following contributions:

- To learn the unknown parameter that linearly parameterizes the reward functions, we instead propose to construct a kind of *composite* feature vectors, weighted by the transition probabilities and the opponent's policy. Intuitively, composite features can be seen as features of corresponding information set-action pairs. Equipped with such composite features, we further propose a "least-squares loss estimator" for this problem, which we call *fictitious* least-squares loss estimator since it is not a true least-squares loss estimator, due to that the "feature covariance matrix" of the fictitious least-squares loss estimator is weighted by the sequence-form policies instead of any probability distributions. Though the fictitious least-squares loss estimator is not a true least-squares loss estimator, we prove that it indeed serves as an unbiased estimator of the unknown reward parameter (see Section 3.1 for details).
- Via integrating our proposed fictitious least-squares loss estimator into the FTRL framework, we propose **F**ictitious least-squares **F**ollow-**T**he-**R**egularized-**L**eader (F^{2}TRL) algorithm. We prove that the regret upper bound of F^{2}TRL is of order $\widetilde{\mathcal{O}}(\lambda\sqrt{dH^2T})$ in large T regime, where d is the ambient dimension of the feature mapping, H is the horizon length, λ is a "balance coefficient" and T is the number of episodes. In particular, λ is moderately large when the environment state transition is nearly a uniform distribution (specifically, $\lambda \leq 1$ when the environment state transition is uniformly at random and the game tree is a k-ary tree). Moreover, we show that λ can only be as large as X in the worst case, guaranteed by the design of our new "balanced transition" over information set-action space, and this worst-case hardly happens in practice (see Section 3 for further details). At the core of both the design and analysis of our F^{2}TRL algorithm is the newly proposed "balanced transition", which might be of independent interest.

- To complement the results of our regret upper bound, we also establish the first regret lower bound of order $\Omega(\sqrt{d\min(d,H)T})$ for learning IIEFGs with linearly parameterized rewards. Moreover, empirical evaluations are conducted on various environments, which corroborate the advantages of our methods against previous ones (see Section 5 for details).

1.1 ADDITIONAL RELATED WORKS

In addition to tabular IIEFGs/POMGs, the other line of research most related to our work is learning fully observable MGs with function approximation [Xie et al., 2020, Chen et al., 2022, Xiong et al., 2022, Jin et al., 2022, Wang et al., 2023, Cui et al., 2023, Ni et al., 2023, Zhang et al., 2023]. These works generally fall into two categories. The first category aims to relax the assumption of linear function approximation by studying MGs with general function approximation [Xiong et al., 2022, Jin et al., 2022, Ni et al., 2023], and the other category of works focuses on learning general-sum MGs [Wang et al., 2023, Cui et al., 2023, Ni et al., 2023, Zhang et al., 2023]. However, we note that all these works study *fully observable* MGs with function approximation, which assume the underlying states are observable to the players and thus are not applicable for solving POMGs. To our knowledge, there are no existing works studying POMGs with function approximation, which is the main focus of this work.

2 PRELIMINARIES

For ease of discussion, we study IIEFGs in the formulation of POMGs [Kozuno et al., 2021, Bai et al., 2022]. In this section, we introduce the preliminaries of POMGs.

Partially Observable Markov Games An episodic, finite-horizon, two-player zero-sum POMG is denoted by $\mathrm{POMG}(H, \mathcal{S}, \mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r)$, in which

- H is the length of the horizon;
- $\mathcal{S} = \bigcup_{h\in[H]} \mathcal{S}_h$, where $\mathcal{S}_h \bigcap \mathcal{S}_{h'} = \emptyset$ for all $h \neq h'$, is a finite state space with cardinality $S = \sum_{h=1}^{H} S_h$ and $|\mathcal{S}_h| = S_h, \forall h \in [H]$;
- $\mathcal{X} = \bigcup_{h\in[H]} \mathcal{X}_h$ is the finite space of information sets (short for *infosets* in what follows) for the max-player, where $\mathcal{X}_h = \{x(s) : s \in \mathcal{S}_h\}$ with $x : \mathcal{S} \to \mathcal{X}$ as the emission function and $\mathcal{X}_h \bigcap \mathcal{X}_{h'} = \emptyset$ for all $h \neq h'$. The cardinality X of $\mathcal{X}$ satisfies $X = \sum_{h=1}^{H} X_h$ with $|\mathcal{X}_h| = X_h$. The finite space of infosets $\mathcal{Y} = \bigcup_{h\in[H]} \mathcal{Y}_h$ for the min-player and associated quantities are defined analogously;
- $\mathcal{A}$ with $|\mathcal{A}| = A$ and $\mathcal{B}$ with $|\mathcal{B}| = B$ are the finite action spaces for the max-player and min-player, respectively;
- $\mathbb{P} = \{p_0(\cdot) \in \Delta_{\mathcal{S}_1}\} \bigcup \{p_h(\cdot|s_h, a_h, b_h) \in \Delta_{\mathcal{S}_{h+1}}\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}, h\in[H-1]}$ are the state transition probabilities, with $p_0(\cdot)$ as the probability distribution over initial states and $p_h(s_{h+1}|s_h, a_h, b_h)$ as the probability of transitioning to the next state s_{h+1} conditioned on (s_h, a_h, b_h) at step h;
- $r = \{r_h(s_h, a_h, b_h) \in [-1, 1]\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}, h\in[H]}$ are the random reward functions with $\bar{r}_h(s_h, a_h, b_h)$ as means.

Learning Protocol Let $\mu = \{\mu_h\}_{h\in[H]}$ be the max-player's (stochastic) policy, where $\mu_h : \mathcal{X}_h \to \Delta_{\mathcal{A}}$. We denote by $\Pi_{\max} = \{\mu : \mathcal{X} \to \Delta_{\mathcal{A}}\}$ the set of the policies of the max-player. Similarly, the min-player's (stochastic) policy is defined as $\nu = \{\nu_h\}_{h\in[H]}$ and the set of the policies of the min-player is denoted by $\Pi_{\min}$. The game proceeds in T episodes. At the beginning of episode t, the max-player and the min-player choose policies $\mu_t \in \Pi_{\max}$ and $\nu_t \in \Pi_{\min}$, respectively. Then, an initial state s_1^t will be sampled from $p_0(\cdot)$. At each step h, the max-player and min-player will only observe their infosets $x_h^t := x(s_h^t)$ and $y_h^t := y(s_h^t)$ respectively, but *without* observing s_h^t. Conditioned on x_h^t, the max-player will take an action $a_h^t \sim \mu_h^t(\cdot|x_h^t)$ and simultaneously the min-player will take an action $b_h^t \sim \nu_h^t(\cdot|y_h^t)$. Subsequently, the game will transition to the next state $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t, b_h^t)$. Meanwhile, the max-player and min-player will receive rewards $r_h^t := r_h(s_h^t, a_h^t, b_h^t)$ and $-r_h^t$ respectively. The t-th episode will terminate after the max-player and the min-player take actions a_H^t and b_H^t and receive rewards r_H^t and $-r_H^t$, respectively.

Perfect Recall and Tree Structure As in previous works [Kozuno et al., 2021, Bai et al., 2022, Fiegel et al., 2023], we suppose that the POMGs satisfy the *tree structure* and the *perfect recall* condition [Kuhn, 1953]. Specifically, the tree structure means that for any $h = 2, \ldots, H$ and $s_h \in \mathcal{S}_h$, there exists a *unique* trajectory $(s_1, a_1, b_1, \ldots, s_{h-1}, a_{h-1}, b_{h-1})$ leading to s_h. Perfect recall condition holds for each player if for any $h = 2, \ldots, H$ and any infoset $x_h \in \mathcal{X}_h$, there exists a *unique* history $(x_1, a_1, \ldots, x_{h-1}, a_{h-1})$ leading to x_h and similarly for the min-player. In addition, we denote by $C_{h'}(x_h, a_h) \subset \mathcal{X}_{h'}$ the descendants of (x_h, a_h) at step $h' \geq h$. With a slight abuse of notation, we also let $C_{h'}(x_h) := \cup_{a_h\in\mathcal{A}} C_{h'}(x_h, a_h)$ and $C(x_h, a_h) := C_{h+1}(x_h, a_h)$.

Sequence-form Representations For any pair of product policy (μ, ν), the tree structure and perfect recall condition enable the *sequence-form representations* of the reaching probability of state-action (s_h, a_h, b_h):

$$\begin{aligned}&\mathbb{P}^{\mu,\nu}(s_h, a_h, b_h)\\ =&p_{1:h}(s_h)\mu_{1:h}(x(s_h), a_h)\nu_{1:h}(y(s_h), b_h)\,,\end{aligned} \tag{1}$$

where $p_{1:h}(s_h) = p_0(s_1)\prod_{h'=1}^{h-1} p_{h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'})$ is the sequence-form transition probability,

$\mu_{1:h}(x_h, a_h) \coloneqq \prod_{h'=1}^{h} \mu_{h'}(a_{h'}|x_{h'})$ and $\nu_{1:h}(y_h, b_h) \coloneqq \prod_{h'=1}^{h} \nu_{h'}(b_{h'}|y_{h'})$ are the sequence-form policies. Under sequence-form representations, we slightly abuse the meanings of μ and ν by viewing $\mu = \{\mu_{1:h}\}_{h\in[H]}$ and $\nu = \{\nu_{1:h}\}_{h\in[H]}$. Also, it is clear that $\Pi_{\max}$ is a convex compact subspace of $\mathbb{R}^{XA}$ satisfying constraints $\mu_{1:h}(x_h, a_h) \geq 0$ and $\sum_{a_h\in\mathcal{A}} \mu_{1:h}(x_h, a_h) = \mu_{1:h-1}(x_{h-1}, a_{h-1})$ with (x_{h-1}, a_{h-1}) being such that $x_h \in C(x_{h-1}, a_{h-1})$ (understanding $\mu_{1:0}(x_0, a_0) = p(\emptyset) = 1$).

In this work, we assume that the player has access to the knowledge of the sequence-form transition probabilities, as explained in the following assumption.

Assumption 1. *The sequence-form transition probability $p_{1:h}(s_h)$ of any s_h is known.*

Remark 1. *Note that this is slightly weaker than assuming knowing $\mathbb{P}$ as Assumption 1 only assumes $p_{1:h}(s_h) = p_0(s_1)\prod_{h'=1}^{h-1} p_{h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'})$ is known. Though this assumption is not required as in previous works studying tabular POMGs [Kozuno et al., 2021, Bai et al., 2022], we remark that a similar assumption of knowing $\mathbb{P}$ is also required by Neu and Olkhovskaya [2021], which initiates the first step for learning adversarial linear MDPs. We leave the question of whether this assumption can be eliminated in our problem as our future work.*

POMGs with Linear Function Approximation We now introduce the definition of linear realizability over the reward functions of POMGs, detailed as follows.

Definition 1 (Linear Rewards in POMGs)**.** *The reward function r in $\mathrm{POMG}(\mathcal{S}, \mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}, H, \mathbb{P}, r)$ is linearly realizable with a known feature mapping $\boldsymbol{\phi} : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \mathbb{R}^d$ if for each $h \in [H]$, there exists an unknown parameter vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$ such that $\bar{r}_h(s_h, a_h, b_h) = \langle \boldsymbol{\phi}(s_h, a_h, b_h), \boldsymbol{\theta}_h\rangle$ for any $(s_h, a_h, b_h) \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{B}$. Further, we assume that $\sup_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}} \|\boldsymbol{\phi}(s_h, a_h, b_h)\|_2 \leq L$ and $\{\boldsymbol{\phi}(s_h, a_h, b_h)\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}}$ spans $\mathbb{R}^d$, $\forall h \in [H]$.*

Similar definitions of linear reward functions can also be seen in fully observable linear MGs [Xie et al., 2020]. However, as we shall see in Section 3.1, the imperfect information in POMGs brings significant difficulties in utilizing the linear structure over the reward functions compared with fully observable MGs. Note that the regularity assumption that the range of $\bar{r}_h(\cdot,\cdot,\cdot)$ and the norm of $\boldsymbol{\phi}(\cdot,\cdot,\cdot)$ are bounded is only for the purpose of normalization, and the assumption that $\mathbb{R}^d$ is spanned by the feature vectors is for convenience only [Lattimore and Szepesvári, 2020].

Learning Objective For any product policy (μ, ν), denote by $V^{\mu,\nu} = \mathbb{E}_{\mu,\nu}\left[\sum_{h=1}^{H} r_h(s_h, a_h, b_h)\right]$ the value function of (μ, ν), where the expectation is taken over the randomness of the policies (μ, ν) and the environment. In this paper, we focus on the learning objective of regret minimization. W.l.o.g., we consider the case where the max-player is the learning agent, and the min-player is the (potentially adversarial) opponent, who might choose her policy ν^t arbitrarily, probably based on all the history information (including the knowledge of $\{\mu^k\}_{k=1}^{t-1}$) up to episode $t-1$. Formally, the max-player aims to design policies $\{\mu^t\}_{t=1}^{T}$ to minimize the *pseudo-regret* (*regret* for short) compared with the best fixed policy $\mu^\dagger$ in hindsight:

$$\mathfrak{R}_{\max}^{T} = \max_{\mu^\dagger\in\Pi_{\max}} \mathbb{E}\left[\sum_{t=1}^{T}\left(V^{\mu^\dagger,\nu^t} - V^{\mu^t,\nu^t}\right)\right], \quad (2)$$

where the expectation is taken over the (potential) randomness of both the max-player and min-player.

Additional Notations We slightly abuse the notation to view x_h as the set $\{s \in \mathcal{S}_h : x(s) = x_h\}$, when writing $s \in x_h$. With sequence-form representations, for any $\mu \in \Pi_{\max}$ and a sequence of functions $f = (f_h)_{h\in[H]}$ with $f_h : \mathcal{X}_h \times \mathcal{A} \to \mathbb{R}$, let $\langle \mu, f\rangle \coloneqq \sum_{h\in[H],(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h, a_h) f_h(x_h, a_h)$. We denote by $\mathcal{F}^t$ the σ-algebra generated by $\{(s_h^k, a_h^k, b_h^k, r_h^k)\}_{h\in[H],k\in[t]}$. For simplicity, we abbreviate $\mathbb{E}[\cdot \mid \mathcal{F}^t]$ as $\mathbb{E}^t[\cdot]$. The notation $\widetilde{\mathcal{O}}(\cdot)$ in this paper suppresses all the logarithmic factors.

3 FICTITIOUS LEAST-SQUARES FOLLOW-THE-REGULARIZED-LEADER

In Section 3.1, we present the proposed *fictitious* least-squares loss estimator for learning the unknown reward parameter. Subsequently, in Section 3.2, we provide the algorithmic details of the F^{2}TRL algorithm, along with its pseudocode shown in Algorithm 1.

3.1 FICTITIOUS LEAST-SQUARES LOSS ESTIMATOR

For a fixed ν^t, Eq. (1) indicates that the value function V^{μ^t,ν^t} is linear in μ^t [Kozuno et al., 2021]:

$$V^{\mu^t,\nu^t} = \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}^t(x_h, a_h) \times \sum_{s_h\in x_h, b_h\in\mathcal{B}} p_{1:h}(s_h)\, \nu_{1:h}^t(y(s_h), b_h)\, \bar{r}_h(s_h, a_h, b_h).$$

Hence, the regret in Eq. (2) can be rewritten as $\mathfrak{R}_{\max}^{T} = \max_{\mu^\dagger\in\Pi_{\max}} \mathbb{E}\left[\sum_{t=1}^{T}\langle \mu^t - \mu^\dagger, \ell^t\rangle\right]$, where ℓ_h^t is *loss*

function in round t such that

$$\ell_h^t(x_h, a_h) := -\sum_{s_h \in x_h, b_h \in \mathcal{B}} p_{1:h}(s_h)\, \nu_{1:h}^t(y(s_h), b_h)\, \bar{r}_h(s_h, a_h, b_h)\,.$$

This implies that one can translate the regret minimization in Eq. (2) into a linear regret minimization problem.

To learn the unknown parameter $\boldsymbol{\theta}_h$ with the leverage of the linear structure over the reward function, one may construct some linear loss estimator $\hat{\boldsymbol{\theta}}_h$ of $\boldsymbol{\theta}_h$. However, this is more challenging in our case than that of linear bandits [Abbasi-Yadkori et al., 2011], linear MDPs [Jin et al., 2020], and fully observable linear MGs [Xie et al., 2020], as the learning agent only observes the infoset $x(s_h)$ and does not even know the underlying state s_h and its associated feature vector $\boldsymbol{\phi}(s_h, a_h, b_h)$, making it impossible to regress $r_h(s_h, a_h, b_h)$ against $\boldsymbol{\phi}(s_h, a_h, b_h)$. To cope with this issue and build a "least-squares loss estimator", we instead consider constructing the following feature vector for each (x_h, a_h), which is a composite feature vector weighted by the opponent's policy ν^t and transition $\mathbb{P}$:

$$\boldsymbol{\phi}^{\nu^t}(x_h, a_h) := -\sum_{(s_h, b_h) \in x_h \times \mathcal{B}} p_{1:h}(s_h)\nu_{1:h}^t(y(s_h), b_h)\boldsymbol{\phi}(s_h, a_h, b_h)\,.$$

Intuitively, the constructed composite feature vector $\boldsymbol{\phi}^{\nu^t}(x_h, a_h)$ can be regarded as the "feature vector" of corresponding infoset-action (x_h, a_h). [1] Further, one can see that $\ell_h^t(x_h, a_h)$ is indeed linear with $\boldsymbol{\theta}_h$ and $\boldsymbol{\phi}^{\nu^t}(x_h, a_h)$:

$$\begin{aligned}&\left\langle -\boldsymbol{\phi}^{\nu^t}(x_h, a_h), \boldsymbol{\theta}_h \right\rangle \\ =&\left\langle \sum_{(s_h, b_h) \in x_h \times \mathcal{B}} p_{1:h}(s_h)\nu_{1:h}^t(y(s_h), b_h)\boldsymbol{\phi}(s_h, a_h, b_h), \boldsymbol{\theta}_h \right\rangle \\ =&-\ell_h^t(x_h, a_h)\,.\end{aligned}$$

Based on $\boldsymbol{\phi}^{\nu^t}(x_h, a_h)$, we define the "feature covariance matrix" $\boldsymbol{Q}_{\mu,h}^t$ for any policy μ at step h in episode t as

$$\boldsymbol{Q}_{\mu,h}^t = \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h)\, \boldsymbol{\phi}^{\nu^t}(x_h, a_h)\boldsymbol{\phi}^{\nu^t}(x_h, a_h)^\top\,. \tag{3}$$

We are now ready to introduce the proposed "least-squares loss estimator" $\hat{\boldsymbol{\theta}}_h^t$:

$$\hat{\boldsymbol{\theta}}_h^t = -(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}\boldsymbol{\phi}^{\nu^t}(x_h^t, a_h^t)r_h^t(s_h^t, a_h^t, b_h^t)\,, \tag{4}$$

which we call the *fictitious* least-squares loss estimator. Importantly, we show that $\hat{\boldsymbol{\theta}}_h^t$ is an unbiased estimator of the unknown $\boldsymbol{\theta}_h$, guaranteed in the following lemma. Its proof is deferred to Appendix A.1.

Lemma 1. *For any $t \in [T]$ and $h \in [H]$, it holds that $\mathbb{E}^{t-1}[\hat{\boldsymbol{\theta}}_h^t] = \boldsymbol{\theta}_h$.*

Remark 2. *Intuitively, this "least-squares loss estimator" shares a similar spirit as its counterpart in adversarial linear bandit literature [Lattimore and Szepesvári, 2020]. However, note that there are two crucial distinctions between $\hat{\boldsymbol{\theta}}_h^t$ defined above and the common least-squares loss estimator in adversarial linear bandits: (a) $\mu_{1:h}^t(\cdot,\cdot)$ in the definition of the "feature covariance matrix" $\boldsymbol{Q}_{\mu,h}^t$ is not necessarily a probability distribution over $\mathcal{X}_h \times \mathcal{A}$ (thus $\boldsymbol{Q}_{\mu,h}^t$ itself is not a true feature covariance matrix); and (b) the "feature vector" $\boldsymbol{\phi}^{\nu^t}(x_h^t, a_h^t)$ is de facto not necessarily linear with the regressand $r_h^t(s_h^t, a_h^t, b_h^t)$ (recall $\bar{r}_h(s_h, a_h, b_h) = \langle \boldsymbol{\phi}(s_h, a_h, b_h), \boldsymbol{\theta}_h \rangle$, which only means $\bar{r}_h$ is linear in $\boldsymbol{\phi}(\cdot,\cdot,\cdot)$ instead of $\boldsymbol{\phi}^{\nu^t}(\cdot,\cdot)$). Due to the above two reasons, $\hat{\boldsymbol{\theta}}_h^t$ is not a real least-squares loss estimator and this is why we term $\hat{\boldsymbol{\theta}}_h^t$ as the fictitious least-squares loss estimator. On the other hand, as shown in Lemma 1, via constructing $\hat{\boldsymbol{\theta}}_h^t$, we indeed address the challenge that we can not regress $r_h(s_h, a_h, b_h)$ against $\boldsymbol{\phi}(s_h, a_h, b_h)$ due to the partial observability in POMGs.*

Remark 3. *When constructing the composite feature vector $\boldsymbol{\phi}^{\nu^t}(x_h, a_h)$, our algorithm uses the product of the sequence-form transition probability $p_{1:h}(s_h)$ and the sequence-form policy $\nu_{1:h}^t(y(s_h), b_h)$ to weight the feature vectors over state-action triplets. Some works studying adversarial linear Markov decision processes (MDPs) (e.g., Kong et al. [2024] and Liu et al. [2024]) use the occupancy measure (OM) $\mu^{\pi,p}(s,a)$, which is the probability of visiting state-action pair (s,a) under policy π and transition probability p, to weight the feature vectors over state-action pairs. We would like to note that there remain several key differences between our idea and theirs. First, from an algorithmic design perspective, for each infoset-action pair (x_h, a_h), our weighting operation is performed only on a subset of state-actions $\{s_h \in \mathcal{S}_h : x(s_h) = x_h\} \times \mathcal{B}$ and the weight $p_{1:h}(\cdot)\nu_{1:h}^t(\cdot,\cdot)$ actually is not a probability measure over $\{s_h \in \mathcal{S}_h : x(s_h) = x_h\} \times \mathcal{B}$. In contrast, such a weighting operation in works of Kong et al. [2024] and Liu et al. [2024] is performed on all the state-action pairs $\mathcal{S}_h \times \mathcal{A}$ ($\mathcal{S}_h$ is the set of all the states on step h of a layered MDP) and $\mu^{\pi,p}(\cdot,\cdot)$ is a probability measure over $\mathcal{S}_h \times \mathcal{A}$. More importantly, the purpose of our weighting operation is mainly to construct a kind of composite feature vector $\phi^{\nu^t}(x_h, a_h)$ for infoset-action pairs (x_h, a_h) so as to construct an unbiased least-squares loss estimator. After such a weighting operation, for each step h, we obtain a set of feature vectors $\{\phi^{\nu^t}(x_h, a_h)\}_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}$. On the contrary, the weighting operation in works of Kong et al.*

[1] Note that our construction of $\boldsymbol{\phi}^{\nu^t}(x_h, a_h)$ depends on the knowledge of the min-player's policy ν_t. While this is not necessary in some works on tabular POMGs [Kozuno et al., 2021, Bai et al., 2022], the requirement for knowledge of opponents' policies—and even the more restrictive assumption that all players are controlled by a central controller—can be seen in various studies on (fully-observable) MGs with linear function approximation (*e.g.*, [Chen et al., 2022, Xie et al., 2020, Cui et al., 2023]).

[2024] and Liu et al. [2024] is to construct a kind of feature vector ϕ^π for each policy π so as to reduce learning adversarial linear MDPs into the problem of learning adversarial linear bandits with $(\phi^\pi)_{\pi\in\Pi}$ as the underlying action set.

Algorithm 1 F^{2}TRL (max-player version)

1: **Input:** Tree-like structure of $\mathcal{X}\times\mathcal{A}$, learning rates η and "balanced transition" $p^\star$.
2: **Initialization:** Set μ^1 as the uniform policy.
3: **for** $t=1$ to T **do**
4: **for** $h=1$ to H **do**
5: Observe infoset x_h^t, execute action $a_h^t\sim\mu_h^t(\cdot|x_h^t)$ and receive reward $r_h^t(s_h^t,a_h^t,b_h^t)$.
6: **end for**
7: Construct composite features $\{\boldsymbol{\phi}^{\nu^t}(x,a)\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$.
8: **for** $h=1$ to H **do**
9: Compute $\boldsymbol{Q}^t_{\mu^t,h}$ as defined in Eq. (3).
10: Compute $\hat{\boldsymbol{\theta}}_h^t$ as defined in Eq. (4).
11: **end for**
12: Construct loss estimate for all (x_h,a_h) and $h\in[H]$: $\hat{\ell}_h^t(x_h,a_h)=\langle\boldsymbol{\phi}^{\nu^t}(x_h,a_h),\hat{\boldsymbol{\theta}}_h^t\rangle$.
13: Compute cumulative loss estimate at episode t: $\hat{L}^t=\hat{L}^{t-1}+\hat{\ell}^t$.
14: Solve Eq. (5) to update policy μ^{t+1} via Algorithm 2.
15: **end for**

3.2 ALGORITHMIC DETAILS

With the constructed fictitious loss estimator in Eq. (4), we are now ready to introduce the algorithmic design of our F^{2}TRL algorithm.

In each episode t, after interacting with the min-player using policy μ^t (Line 4 - Line 6), F^{2}TRL will construct the composite feature vectors and fictitious least-squares loss estimator $\hat{\boldsymbol{\theta}}_h^t$ defined in Eq. (4) (Line 7 - Line 11). Then F^{2}TRL will compute the loss estimate $\hat{\ell}_h^t(x_h,a_h)$ with the leverage of $\hat{\boldsymbol{\theta}}_h^t$ as well as the composite feature vectors for all $(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}$ and $h\in[H]$ and update the cumulative loss estimate $\hat{L}^t$ (Line 12 - Line 13). At the end of episode t, to update the policy μ^{t+1} used in episode $t+1$, it solves the following linear optimization problem regularized by potential function $\{\Psi_h\}_{h\in[H]}$ (Line 14):

$$\mu^{t+1}=\arg\min_{\mu\in\Pi_{\max}}\left\langle\mu,\hat{L}^t\right\rangle+\frac{1}{\eta}\sum_{h=1}^{H}\Psi_h\left(p^\star_{1:h}\cdot\mu_{1:h}\right), \quad (5)$$

where $\hat{L}^t=\sum_{k=1}^t\hat{\ell}^k$ is the cumulative loss estimate, $\Psi_h(w_h)=\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}w_h(x_h,a_h)\log(w_h(x_h,a_h))$ is the negentropy potential function, $p^\star_{1:h}(x_h)=p^\star_0(x_1)\prod_{h'=1}^{h-1}p^\star_{h'}(x_{h'+1}|x_{h'},a_{h'})$ with $p^\star_h(\cdot|x_h,a_h)\in\Delta_{C(x_h,a_h)}$ being a kind of transition probability over $\mathcal{X}_h\times\mathcal{A}\times\mathcal{X}_{h+1}$, and $p^\star_{1:h}\cdot\mu_{1:h}$ is defined as $[p^\star_{1:h}\cdot\mu_{1:h}](x_h,a_h)=p^\star_{1:h}(x_h)\mu_{1:h}(x_h,a_h)$. We note that $p^\star$ is well-defined due to the perfect recall condition, and $p^\star_{1:h}\cdot\mu_{1:h}$ is a probability distribution over the infoset-action space $\mathcal{X}_h\times\mathcal{A}$ at step h.

When bounding the regret of the FTRL, it is essential to make the stability term well-controlled in the analysis. To this end, we construct the following "balanced transition" as our transition probability $p^\star_{1:h}(\cdot)$ over infoset-action space:

$$p^\star=\arg\max_{\tilde{p}\in\mathbb{P}^\star}\min_{h\in[H],x_h\in\mathcal{X}_h}\tilde{p}_{1:h}(x_h), \quad (6)$$

where $\mathbb{P}^\star$ denotes the set of all the valid transition probabilities over infoset-action space. We also remark that similar approaches that utilize FTRL or OMD with "balanced transition" over $\mathcal{X}_h\times\mathcal{A}\times\mathcal{X}_{h+1}$ have also been exploited in previous works (see, *e.g.*, Bai et al. [2022], Fiegel et al. [2023]). However, the design of our "balanced transition" $p^\star_{1:h}(\cdot)$ differs from the previous ones in the following two aspects:

- $p^\star(x_{h+1}|x_h,a_h)$ in this work is proportional to all the reachable infosets $C_H(x_{h+1},a_{h+1})$ in $\mathcal{X}_H$ by taking some fixed action $a_{h+1}\in\mathcal{A}$ at infoset x_{h+1}. In contrast, the "balanced transitions" of Bai et al. [2022], Fiegel et al. [2023] are devised by only considering the reachable infosets in $\mathcal{X}_{h'}$ for some $h'\geq h+1$ or all the reachable infosets in the whole sub-tree.
- Our $p^\star(x_{h+1}|x_h,a_h)$ is contributed by some fixed action $a_{h+1}\in\mathcal{A}$ that maximizes the number of the reachable infosets $|C_H(x_{h+1},a_{h+1})|$ in $\mathcal{X}_H$, while previous "balanced transitions" of Bai et al. [2022], Fiegel et al. [2023] are contributed by the sum of all the reachable infosets by taking all actions $a_{h+1}\in\mathcal{A}$ at infoset x_{h+1}.

As we shall see in Section 4, the property of our $p^\star$ plays a crucial role when bounding the stability term of F^{2}TRL algorithm in the regret analysis.

Computation We prove that the computation of Eq. (5) has a closed-form update and can be solved by backward dynamic programming, as illustrated in Algorithm 2 in Appendix C.1. Besides, the computation of our "balanced transition" $p^\star$ in Eq. (6) can be also efficiently solved by Algorithm 3 and we defer the details to Appendix C.2.

4 ANALYSIS

In this section, we first derive the regret upper bound for our F^{2}TRL algorithm. Then, in Section 4.2, we provide the regret lower bound for learning IIEFGs with linearly parameterized rewards.

4.1 REGRET UPPER BOUND

Let $p^\nu_{1:h}(x_h):=\sum_{s_h\in x_h}p_{1:h}(s_h)\,\nu_{1:h-1}(y(s_{h-1}),b_{h-1})$, which can be seen as the probability of reaching x_h

contributed by environment transition $\mathbb{P} = \{p_h\}_{h=0}^{H-1}$ and opponent's policy ν. Denote by $\beta_h^\star := \min_{x_h \in \mathcal{X}_h} p_{1:h}^\star(x_h)$ and $\beta_h^\nu := \max_{t\in[T], x_h \in \mathcal{X}_h} p_{1:h}^{\nu^t}(x_h)$. Then we define the "balance coefficient" λ as $\lambda := \max_{h\in[H]} \beta_h^\nu / \beta_h^\star$. Besides, let $\rho = \min_{t\in[T], h\in[H]} \rho_{\min}(\boldsymbol{Q}_{\mu^t,h}^t)$ be the minimum of all the minimal eigenvalues of the feature covariance matrices.[2]

We are now ready to present the regret upper bound of F^{2}TRL, the proof of which is postponed to Appendix B.

Theorem 1. *For IIEFGs with linearly realizable rewards and known sequence-form transition probabilities, by setting learning rate $\eta = \sqrt{\frac{2\log(XA)}{Td}}$, the regret of F^{2}TRL is upper bounded by $\mathfrak{R}_{\max}^T \le \mathcal{O}(\exp(L^2\sqrt{\log(XA)}/(\beta_H^\star \rho\sqrt{Td}))\lambda H\sqrt{dT\log(XA)})$.*

Remark 4. *Intuitively, λ measures the balance effect of the "balanced transition" $p_{1:h}^\star$ compared with the transition over infoset-action space contributed by the environment state transition $\mathbb{P}$ and the opponent's policy ν^t. Indeed, due to the design of our "balanced transition" $p^\star$, λ is moderately large when the environment state transition $\mathbb{P}$ is nearly uniform (in particular, $\lambda \le 1$ when the environment state transition $\mathbb{P}$ is a uniform distribution and the game tree is a k-ary tree; see Lemma 9 in Appendix B.3). On the other hand, the design of our "balanced transition" $p^\star$ also guarantees that $\beta_H^\star \ge 1/X$ and thus $\lambda \le X$ in the worst-case scenario (see Lemma 10 in Appendix B.3). Nevertheless, we should note that this worst case is very unlikely to happen in practice unless it simultaneously happens that (a) the environment state transitions along the trajectory $\{(s_h, a_h, b_h)\}_{h\in[H-1]}$ leading to s_H s.t. $p_{1:H}^\star(x(s_H)) = \beta_H^\star$ satisfy $p_h(s_{h+1}|s_h, a_h, b_h) = 1$ for all (s_h, a_h, b_h) along the trajectory; and (b) the opponent knows the underlying environment transitions and the mapping $y : \mathcal{S} \to \mathcal{Y}$ so that the opponent can intentionally ensure $\nu_{1:H-1}^t(y(s_{H-1}), b_{H-1}) = 1$ by setting $\nu^t(b_h|y(s_h)) = 1$ for all (s_h, b_h) along the trajectory. Notice that condition (b) hardly happens in the self-play setting where the policies of the min-player are also generated by an algorithm. Also, if the opponent is a pure adversary aiming to maximize the regret of the max-player and only condition (a) holds, the best that the opponent can do is to uniformly pick an action $b_y \in \mathcal{B}$ at each infoset $y \in \mathcal{Y}$ and set her policy ν^t such that $\nu^t(b_y|y) = 1$. This can only guarantee that $\nu_{1:H-1}^t(y(s_{H-1}), b_{H-1}) = 1$ (and thus $\lambda = X$) happens with an exponentially small probability of $B^{-(H-1)}$. Additionally, we remark that λ has nothing to do with the commonly discussed "concentrability coefficient" in offline RL literature [Kumar et al., 2019], which might be arbitrarily large in practice [Liu et al., 2020, Xie et al., 2021].*

[2] Note that it is guaranteed that $\rho > 0$ due to that μ^1 is set as the uniform policy in Line 2 of Algorithm 1 and the closed-form update of μ^t as shown in Algorithm 2.

Remark 5. *For the adversarial linear bandit problem, a more tractable special case of IIEFGs with linearly parameterized rewards, one can eliminate the dependence of the regret bound on $1/\rho$ by mixing μ^t with an optimal design distribution. However, for general linear IIEFG problems, it is highly unclear how to achieve this, as for all $\mu \in \Pi_{\max}$, $\mu_{1:h}(\cdot,\cdot)$ is not even a valid probability distribution over $\mathcal{X}_h \times \mathcal{A}$. On the other hand, it is worth noting that the dependence on $1/\rho$ only appears in the exponential term of our regret upper bound, which is inversely related to the number of episodes T and approaches 1 as T grows large (i.e., $T \ge \Omega(L^4\log(XA)/((\beta_H^\star\rho)^2 d))$). Besides, we remark that the $\sqrt{\log(XA)}$ dependence has also appeared in previous works studying the more tractable (fully observable) adversarial linear (mixture) MDPs [Neu and Olkhovskaya, 2021, Zhao et al., 2023, Li et al., 2024a,b].*

4.1.1 Technique Overview

In the following, we briefly explain the technical challenges involved in deriving the regret upper bound in Theorem 1 and the approaches we use to address them.

Loss Estimates with Large Negative Magnitudes We bound the regret of F^{2}TRL by considering the common analysis scheme of FTRL to decompose the regret into the *penalty* term and the *stability* term. However, when bounding the stability term, simply following the previous analysis for tabular IIEFGs and other online learning problems with linear function approximation (say, adversarial linear bandits) does not address our problem. In detail, we note that the analysis of Fiegel et al. [2023] to the bound the stability term can only work with non-negative loss estimates, which naturally hold in the tabular case but not in the linear case. A plausible remedy from the analysis of adversarial linear bandits (see, *e.g.*, Chap. 27 of Lattimore and Szepesvári [2020]) is to explicitly bound the Bregman divergence between $\nabla\Psi(\mu^t)$ and $\nabla\Psi(\mu^t) - \eta\hat{\ell}^t$ and use the inequality $\exp(x) \le 1 + x + x^2$ for $x \le 1$, which in turn requires $\eta\hat{\ell}_h^t(x_h, a_h)/p_{1:h}^\star(x_h, a_h) \ge -1$ in our case. Unfortunately, $\eta\hat{\ell}_h^t(x_h, a_h)/p_{1:h}^\star(x_h, a_h) \ge -1$ does not hold in our case and thus thwarts this remedy. To tackle this issue, we instead first evaluate the Bregman divergence between $\nabla\Psi(\mu^t)$ and $\nabla\Psi(\mu^t) - \eta\hat{\ell}^t$ using a local norm regarding some $z^t = \alpha\mu^t + (1-\alpha)\tilde{\mu}^{t+1}$ for some $\alpha \in [0,1]$, where $\tilde{\mu}^{t+1} := \nabla\Psi^\star(\nabla\Psi(\mu^t) - \eta\hat{\ell}^t)$ and $\Psi^\star$ is the convex conjugate of Ψ. Then, with the observation that $z_{1:h}^t(x_h, a_h) \le \max\{\mu_{1:h}^t(x_h, a_h), \tilde{\mu}_{1:h}^{t+1}(x_h, a_h)\}$ and $\tilde{\mu}_{1:h}^{t+1}(x_h, a_h)$ is proportional to $\mu_{1:h}^t(x_h, a_h)$, we bound this local norm by upper bounding $z_{1:h}^t(x_h, a_h)$ using $\mu_{1:h}^t(x_h, a_h)$ (see Lemma 6 for details).

Non-zero Loss Estimates and Balanced Effects of $p^\star$ Moreover, in IIEFGs, bounding the stability term critically

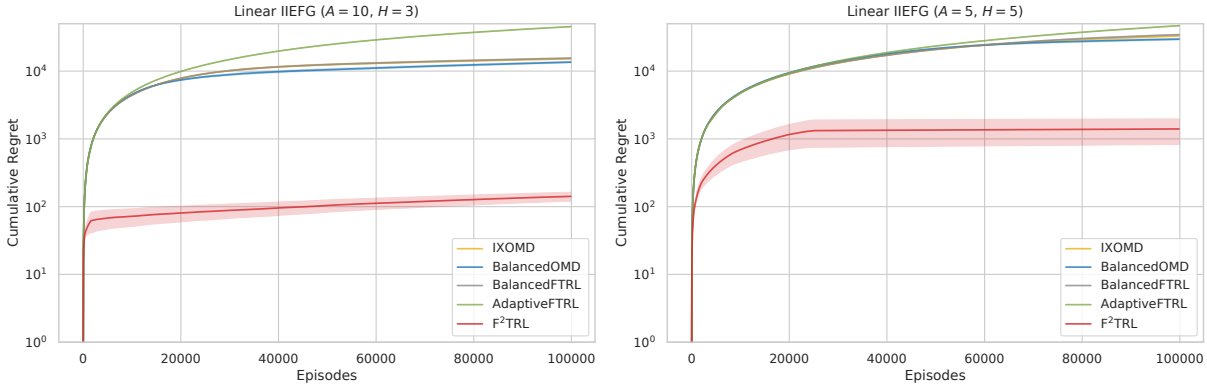

Figure 1: Experimental results of baseline methods and our F²TRL algorithm on two linear IIEFG environments. The curves depict the value of Eq. (2) as a function of the number of episodes, averaged over 10 different seeds, with shaded areas representing the 1 standard error.

relies on the closed-form update of the policies (*e.g.*, Eq. (5) in our case). Nevertheless, previous works on tabular IIEFGs [Kozuno et al., 2021, Bai et al., 2022, Fiegel et al., 2023] solve the updates similar to Eq. (5) by heavily relying on the sparsity of the importance-weighted loss estimator. Specifically, only the infoset-action pairs along the experienced trajectory $\{(x_h^t, a_h^t)\}_{h\in[H]}$ have non-zero loss estimates, while all other infoset-action pairs have zero loss estimates in the tabular case. However, in our linear case, all infoset-action pairs can have non-zero loss estimates, making the methods in previous works inapplicable. We address this challenge by recursively considering all descendants of a given infoset-action pair (x_h, a_h), rather than just one descendant (x_{h+1}^t, a_{h+1}^t) from the experienced trajectory in episode t, when updating the policy $\mu_h^{t+1}(a_h|x_h)$ (see Appendix C.1 for details). Finally, deriving the overall bound for the stability term in our case requires particular care to bound the ratio $\beta_h^\nu/(\min_{x_h} \tilde{p}_{1:h}(x_h))$ for any "balanced transition" $\tilde{p}$ adopted by the algorithm. To make this ratio well-controlled, we seek to maximize $\min_{x_h} \tilde{p}_{1:h}(x_h)$ for all $h \in [H]$. This is exactly facilitated by the design of our "balanced transition" $p^\star$ defined in Eq. (6) computed in Algorithm 3, which guarantees that this ratio is upper bounded by the desired "balance coefficient" λ (please refer to Lemma 7 for details).

4.2 REGRET LOWER BOUND

The following theorem provides the regret lower bound of learning IIEFGs with linearly realizable rewards and known state transition probabilities, the proof of which is deferred to Appendix D.

Theorem 2. *Suppose $A \geq 2$, $d \geq 2$ and $T \geq 2d^2$. Then for any algorithm* Alg *that controls the max-player, generates and executes policies $\{\mu^t\}_{t\in[T]}$, there exists an IIEFG instance with linearly realizable rewards and known state transition probabilities on which $\mathfrak{R}_{\max}^T \geq \Omega(\sqrt{d\min(d,H)T})$.*

Remark 6. *Note that both the regret upper and lower bounds of our algorithm do not have polynomial dependence on X and A, as opposed to the $\Omega(\sqrt{XAT})$ regret lower bound of Bai et al. [2022]. However, we would like to note that this does not imply that our results contradict those of previous works, as both our regret upper and lower bounds are specifically established for IIEFG instances with linear structures over reward functions, while the regret lower bound of Bai et al. [2022] is derived by considering learning on the IIEFG instances without any function approximation structures (i.e., tabular rewards). Besides, we conjecture that the lower bound might be further improved to $\mathfrak{R}_{\max}^T \geq \Omega(\sqrt{dHT})$, and currently the regret upper bound of F²TRL is loose by an $\widetilde{\mathcal{O}}(\sqrt{H})$ factor in large T regime (omitting dependence on λ). The investigation into the possible improvements of the upper and lower bounds is an interesting and also challenging future direction, and we leave this extension as our future study.*

5 EXPERIMENTS

This section presents the empirical evaluations of our F²TRL algorithm as well as previous methods. [3]

Environments We construct two A-ary tree IIEFG environments with linear structures, both of which exactly follow from the construction of the hard-to-learn IIEFG instances used to prove the regret lower bound (please see Appendix D for details of such instances). The IIEFG instance in the first environment involves $H = 3$ steps and $A = 10$ actions at each infoset of the max-player (hence there are 1110 infoset-action pairs of the max-player in total), while the

[3] Codes of the experiments are available at https://github.com/AnonymousXX-XX/Linear-IIEFG.

second IIEFG instance has $H = 5$ steps and $A = 5$ actions at each infoset of the max-player (hence 3905 infoset-action pairs in total). In both environments, the rewards for all state-action pairs $(s, a) \in \bigcup_{h\in[H-1]} \mathcal{S}_h \times \mathcal{A}$ are set to be 0 and the mean of the reward for each $(s, a) \in \mathcal{S}_H \times \mathcal{A}$ is set as $\bar{r}_H(s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta} \rangle$. Particularly, the feature $\boldsymbol{\phi}(s, a)$ has dimension $d = 10$, with each dimension first uniformly sampled from $[-1, 1]$ and then normalized by its L^2-norm, and the construction of $\boldsymbol{\theta}$ is given by the same procedure.

Baselines We incorporate the algorithms in most related works as baselines, including `IXOMD` [Kozuno et al., 2021], `BalancedOMD` [Bai et al., 2022], and `BalancedFTRL`, `AdaptiveFTRL` [Fiegel et al., 2023]. [4] Following Fiegel et al. [2023], we conduct a (logarithmic) grid search on the learning rates of each algorithm in each environment.

Results As shown in Figure 1, the baseline methods except `AdaptiveFTRL` have similar performance in both environments and `AdaptiveFTRL` converges relatively slower than other baselines. Further, `F`^{2}`TRL` outperforms all the baselines with significantly faster convergence rates on both environments, due to the leverage of the linear structures of the games. Besides, all the algorithms empirically suffer more regret in the second environment than in the first one, since it involves a longer horizon length H and more infoset-action pairs to learn than the first environment.

6 CONCLUSIONS

In this work, we make the first step towards provably efficient learning of two-player zero-sum IIEFGs with linear function approximation, in the formulation of POMGs with linearly realizable rewards. It is proven that the proposed `F`^{2}`TRL` algorithm attains a regret guarantee of order $\widetilde{\mathcal{O}}(\lambda H\sqrt{dT})$ in large T regime. We accomplish this by devising a *fictitious* least-squares loss estimator for this problem, along with the design of a kind of new "balanced transition" over infoset-action space, which might be of independent interest. Moreover, we establish an $\Omega(\sqrt{d\min(d, H)T})$ regret lower bound for this problem and conduct empirical evaluations on various environments, which validate the advantages of our `F`^{2}`TRL` algorithm. Besides, there are also several interesting future directions to be explored. One natural question may be how to obtain high-probability results for this challenging problem so as to find an approximate NE with high-probability. We believe it is possible to extend our results to high-probability ones using self-concordant barrier potential functions and increasing learning rates [Lee et al., 2020]. The other question might be whether it is possible to generalize the proposed algorithm to multi-player general-sum IIEFGs. We believe the results of this work will shed light on better understandings of learning large-scale IIEFGs and we leave these extensions as our further studies.

[4] We adopt the codes of all the baselines implemented by Fiegel et al. [2023]: `https://github.com/anon17893/IIG-tree-adaptation`.

ACKNOWLEDGEMENTS

The corresponding author Shuai Li is partly supported by the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*, 2011.

Yu Bai, Chi Jin, Song Mei, and Tiancheng Yu. Near-optimal learning of extensive-form games with imperfect information. In *International Conference on Machine Learning, ICML 2022*. PMLR, 2022.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 2018.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*. PMLR, 2019.

Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting CFR+ and alternating updates. *J. Artif. Intell. Res.*, 2019.

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

Qiwen Cui, Kaiqing Zhang, and Simon S. Du. Breaking the curse of multiagents in a large state space: RL in markov games with independent linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*. PMLR, 2023.

Gabriele Farina and Tuomas Sandholm. Model-free online learning in unknown sequential decision making problems and games. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 2021.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Stochastic regret minimization in extensive-form games. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, 2020.

Gabriele Farina, Robin Schmucker, and Tuomas Sandholm. Bandit linear optimization for sequential decision making and extensive-form games. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 2021.

Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Rémi Munos, Vianney Perchet, and Michal Valko. Adapting to game trees in zero-sum imperfect information games. In *International Conference on Machine Learning, ICML 2023*. PMLR, 2023.

Haobo Fu, Weiming Liu, Shuang Wu, Yijia Wang, Tao Yang, Kai Li, Junliang Xing, Bin Li, Bo Ma, Qiang Fu, and Wei Yang. Actor-critic policy optimization in a large-scale imperfect-information game. In *The Tenth International Conference on Learning Representations, ICLR 2022*, 2022.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, 2015.

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Math. Oper. Res.*, 2010.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory, COLT 2020*. PMLR, 2020.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent RL in large state spaces. In *International Conference on Machine Learning, ICML 2022*. PMLR, 2022.

Michael Johanson, Nolan Bard, Marc Lanctot, Richard G. Gibson, and Michael Bowling. Efficient nash equilibrium approximation through monte carlo counterfactual regret minimization. In *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2012*, 2012.

Daphne Koller and Nimrod Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and economic behavior*, 1992.

Daphne Koller, Nimrod Megiddo, and Bernhard Von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and economic behavior*, 1996.

Fang Kong, Xiangcheng Zhang, Baoxiang Wang, and Shuai Li. Improved regret bounds for linear adversarial mdps via linear optimization. *Trans. Mach. Learn. Res.*, 2024, 2024.

Tadashi Kozuno, Pierre Ménard, Rémi Munos, and Michal Valko. Learning in two-player zero-sum partially observable markov games with perfect recall. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 11987–11998, 2021.

Christian Kroer, Kevin Waugh, Fatma Kilinç-Karzan, and Tuomas Sandholm. Faster first-order methods for extensive-form game solving. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC 2015*, pages 817–834. ACM, 2015.

Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Solving large sequential games with the excessive gap technique. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, 2018.

HW Kuhn. Extensive games and the problem of information. *Contributions to the Theory of Games*, 1953.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 2019.

Moyuru Kurita and Kunihito Hoki. Method for constructing artificial intelligence player with abstractions to markov decision processes in multiplayer game of mahjong. *IEEE Trans. Games*, 2021.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael H. Bowling. Monte carlo sampling for regret minimization in extensive games. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009*, 2009.

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

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 2021.

Junjie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, and Hsiao-Wuen Hon. Suphx: Mastering mahjong with deep reinforcement learning. abs/2003.13590, 2020.

Long-Fei Li, Peng Zhao, and Zhi-Hua Zhou. Dynamic regret of adversarial mdps with unknown transition and linear function approximation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, 2024a.

Long-Fei Li, Peng Zhao, and Zhi-Hua Zhou. Improved algorithm for adversarial linear mixture mdps with bandit feedback and unknown transition. In *International Conference on Artificial Intelligence and Statistics, 2024*. PMLR, 2024b.

Haolin Liu, Chen-Yu Wei, and Julian Zimmert. Towards optimal regret in adversarial linear mdps with bandit feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Weiming Liu, Huacong Jiang, Bin Li, and Houqiang Li. Equivalence analysis between counterfactual regret minimization and online mirror descent. In *International Conference on Machine Learning, ICML 2022*. PMLR, 2022.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 2017.

Rémi Munos, Julien Pérolat, Jean-Baptiste Lespiau, Mark Rowland, Bart De Vylder, Marc Lanctot, Finbarr Timbers, Daniel Hennes, Shayegan Omidshafiei, Audrunas Gruslys, Mohammad Gheshlaghi Azar, Edward Lockhart, and Karl Tuyls. Fast computation of nash equilibria in imperfect information games. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*. PMLR, 2020.

John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 1950.

Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, 2021.

Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Zihan Ding, Chi Jin, and Mengdi Wang. Representation learning for low-rank general-sum markov games. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

Martin Schmid, Neil Burch, Marc Lanctot, Matej Moravcik, Rudolf Kadlec, and Michael Bowling. Variance reduction in monte carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, 2019.

Martin Schmid, Matej Moravcik, Neil Burch, Rudolf Kadlec, Joshua Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, Zach Holland, Elnaz Davoodi, Alden Christianson, and Michael Bowling. Player of games. abs/2112.03178, 2021.

Oskari Tammelin. Solving large imperfect information games using CFR+. abs/1407.5042, 2014.

Yuandong Tian, Qucheng Gong, and Yu Jiang. Joint policy search for multi-agent collaboration with imperfect information. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 1996.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent RL with function approximation. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*. PMLR, 2023.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory, COLT 2020*. PMLR, 2020.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, 2021.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang. A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning, ICML 2022*. PMLR, 2022.

Yuheng Zhang, Yu Bai, and Nan Jiang. Offline learning in markov games with general function approximation. In *International Conference on Machine Learning, ICML 2023*. PMLR, 2023.

Canzhe Zhao, Ruofeng Yang, Baoxiang Wang, and Shuai Li. Learning adversarial linear mixture markov decision processes with bandit feedback and unknown transition. In *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory, COLT 2021*. PMLR, 2021.

Yuan Zhou. Lecture 14: Lower bounds for linear bandits. In *IE 498: Online Learning and Decision Making, Fall 2019*, 2019. URL https://yuanz.web.illinois.edu/teaching/IE498fa19/lec_14.pdf.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Martin Zinkevich, Michael Johanson, Michael H. Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20, 2007*, 2007.

SUPPLEMENTARY MATERIAL

A PROPERTIES OF THE FICTITIOUS LEAST-SQUARES LOSS ESTIMATOR

This section presents the proofs of two key properties of the proposed fictitious least-squares loss estimator.

A.1 UNBIASNESS OF THE FICTITIOUS LEAST-SQUARES LOSS ESTIMATOR

Proof of Lemma 1. The definition of $\hat{\boldsymbol{\theta}}_h^t$ in Eq. (4) implies that

$$\begin{aligned}
\mathbb{E}^{t-1}\left[\hat{\boldsymbol{\theta}}_h^t\right] &= \mathbb{E}^{\mu^t,\nu^t}\left[\hat{\boldsymbol{\theta}}_h^t\right] \\
&= \mathbb{E}^{\mu^t,\nu^t}\left[-(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}\cdot\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)\cdot r_h^t(s_h^t,a_h^t,b_h^t)\right] \\
&= \mathbb{E}^{\mu^t,\nu^t}\left[-(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}\cdot\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)\cdot \bar{r}_h(s_h^t,a_h^t,b_h^t)\right] \\
&= -(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}\sum_{x_h\in\mathcal{X}_h}\sum_{s_h\in x_h}\sum_{a_h\in\mathcal{A}}\sum_{b_h\in\mathcal{B}}\mathbb{P}^{\mu^t,\nu^t}(s_h,a_h,b_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\bar{r}_h(s_h,a_h,b_h) \\
&= -(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}\sum_{x_h\in\mathcal{X}_h}\sum_{a_h\in\mathcal{A}}\mu_{1:h}^t(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\sum_{s_h\in x_h}\sum_{b_h\in\mathcal{B}}p_{1:h}(s_h)\nu_{1:h}^t(y(s_h),b_h)\bar{r}_h(s_h,a_h,b_h) \\
&= (\boldsymbol{Q}_{\mu^t,h}^t)^{-1}\sum_{x_h\in\mathcal{X}_h}\sum_{a_h\in\mathcal{A}}\mu_{1:h}^t(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\left\langle\boldsymbol{\phi}^{\nu^t}(x_h,a_h),\boldsymbol{\theta}_h\right\rangle \\
&= (\boldsymbol{Q}_{\mu^t,h}^t)^{-1}\left(\sum_{x_h\in\mathcal{X}_h}\sum_{a_h\in\mathcal{A}}\mu_{1:h}^t(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\right)\boldsymbol{\theta}_h \\
&= \boldsymbol{\theta}_h\,,
\end{aligned}$$

which concludes the proof. □

A.2 VARIANCE OF THE FICTITIOUS LEAST-SQUARES LOSS ESTIMATOR

The following lemma shows that the "variance" of the proposed loss estimator is well controlled.

Lemma 2. *For any $h\in[H]$, it holds that*

$$\mathbb{E}^{t-1}\left[\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu_{1:h}^t(x_h,a_h)\hat{\ell}_h^t(x_h,a_h)^2\right]\le d\,. \tag{7}$$

Proof. It is clear that

$$
\begin{aligned}
&\mathbb{E}^{t-1}\left[\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\hat{\ell}^t_h(x_h,a_h)^2\right]\\
=&\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\mathbb{E}^{\mu^t,\nu^t}\left[\hat{\boldsymbol{\theta}}^t_h(\hat{\boldsymbol{\theta}}^t_h)^\top\right]\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\\
\overset{(i)}{=}&\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\\
&\quad\cdot\mathbb{E}^{\mu^t,\nu^t}\left[r^t_h(s^t_h,a^t_h,b^t_h)^2(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x^t_h,a^t_h)\boldsymbol{\phi}^{\nu^t}(x^t_h,a^t_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\right]\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\\
\overset{(ii)}{\le}&\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\mathbb{E}^{\mu^t,\nu^t}\left[(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x^t_h,a^t_h)\boldsymbol{\phi}^{\nu^t}(x^t_h,a^t_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\right]\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\\
=&\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\\
&\quad\cdot\left(\sum_{(x'_h,a'_h)\in\mathcal{X}_h\times\mathcal{A}}p^{\nu^t}_{1:h}(x'_h)\mu^t_{1:h}(x'_h,a'_h)\boldsymbol{\phi}^{\nu^t}(x'_h,a'_h)\boldsymbol{\phi}^{\nu^t}(x'_h,a'_h)^\top\right)(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\\
=&\operatorname{tr}\left[\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\right)\right.\\
&\quad\left.\cdot\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}p^{\nu^t}_{1:h}(x_h)\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\right)\right]\\
=&\operatorname{tr}\left(\boldsymbol{I}_d\cdot\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}p^{\nu^t}_{1:h}(x_h)\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\right)\\
\overset{(iii)}{\le}&\operatorname{tr}\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\right)\\
=&\operatorname{tr}(\boldsymbol{I}_d)=d\,,
\end{aligned}
$$

where (i) is due to the definition of $\hat{\boldsymbol{\theta}}^t_h$ in Eq. (4); (ii) is by $|r^t_h(s_h,a_h,b_h)|\le 1$ for any $(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}$; and (iii) follows from $p^{\nu^t}_{1:h}(x_h)\le 1$ for any $x_h\in\mathcal{X}_h$. The proof is thus completed. □

B PROOF OF REGRET UPPER BOUND

To start with, notice that $\Pi_{\max}$ is an affine subspace of $\mathbb{R}^{XA}_{\ge 0}$ satisfying X linear constraints: for any $x_h\in\mathcal{X}$,

$$\sum_{a_h\in\mathcal{A}}\mu_{1:h}(x_h,a_h)=\mu_{1:h-1}(x_{h-1},a_{h-1})\,,$$

where (x_{h-1},a_{h-1}) is the unique predecessor of x_h under perfect recall condition. Thus $\Pi_{\max}$ can be decomposed as $\Pi_{\max}=(F+u)\cap\mathbb{R}^{XA}_{\ge 0}$ where F is a linear subspace and $u\in\Pi_{\max}$.

With slight abuse of notations, we further denote $\Psi_\eta(\mu)=\frac{1}{\eta}\sum_{h=1}^H\Psi_h\left(p^\star_{1:h}\cdot\mu_{1:h}\right)$ and define its convex conjugate function $\Psi^\star_\eta$ on $\mathbb{R}^{XA}_{\ge 0}$ as

$$\Psi^\star_\eta(\boldsymbol{y})\coloneqq\sup_{\boldsymbol{x}\in\mathbb{R}^{AX}_{\ge 0}}\langle\boldsymbol{x},\boldsymbol{y}\rangle-\Psi_\eta(\boldsymbol{x})\,. \tag{8}$$

Also, we denote $D_{\Psi^\star_\eta}(\boldsymbol{x},\boldsymbol{y})=\Psi^\star_\eta(\boldsymbol{x})-\Psi^\star_\eta(\boldsymbol{y})-\left\langle\nabla\Psi^\star_\eta(\boldsymbol{y}),\boldsymbol{x}-\boldsymbol{y}\right\rangle$ as the Bregman divergence induced by $\Psi^\star_\eta$. The following lemma shows the canonical regret decomposition of the FTRL framework. [Zimmert and Seldin, 2019, Lattimore and Szepesvári, 2020].

Lemma 3. *The regret of* F^{2}TRL *can be decomposed as*

$$\mathfrak{R}_{\max}^T \leq \underbrace{\max_{\mu\in\Pi_{\max}} [-\Psi_\eta(\mu)]}_{\textsc{Penalty}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T D_{\Psi_\eta^\star}(\nabla\Psi_\eta(\mu^t) - \hat{\ell}^t, \nabla\Psi_\eta(\mu^t))\right]}_{\textsc{Stability}}.$$

Proof. Let $\mu^\dagger \in \Pi_{\max}$ be some policy (in the sequence-form representation). For all $t \in [T]$, the instantaneous regret against $\mu^\dagger$ at step t can be decomposed into

$$\left\langle \mu^t - \mu^\dagger, \hat{\ell}^t \right\rangle = \left[\Phi_\eta\left(-\hat{L}^{t-1}\right) - \Phi_\eta\left(-\hat{L}^t\right) - \left\langle \mu^\dagger, \hat{\ell}^t\right\rangle\right] + \left[\left\langle \mu^t, \hat{\ell}^t\right\rangle + \Phi_\eta\left(-\hat{L}^t\right) - \Phi_\eta\left(-\hat{L}^{t-1}\right)\right],$$

where $\Phi_\eta(\boldsymbol{y}) \coloneqq \sup_{\boldsymbol{\mu}\in\Pi_{\max}} \langle \boldsymbol{\mu}, \boldsymbol{y}\rangle - \Psi_\eta(\boldsymbol{\mu})$.

Taking summation of the above display over t yields

$$\begin{aligned}
&\sum_{t=1}^T \left[\Phi_\eta\left(-\hat{L}^{t-1}\right) - \Phi_\eta\left(-\hat{L}^t\right) - \left\langle \mu^\dagger, \hat{\ell}^t\right\rangle\right] \\
=&\Phi_\eta(0) - \Phi_\eta\left(-\hat{L}^t\right) - \left\langle \mu^\dagger, \hat{L}^t\right\rangle \\
\overset{(i)}{\leq}& \max_{\mu\in\Pi_{\max}} \left[-\Psi_\eta(\mu)\right] + \Psi_\eta\left(\mu^\dagger\right) \\
\overset{(ii)}{\leq}& \max_{\mu\in\Pi_{\max}} \left[-\Psi_\eta(\mu)\right],
\end{aligned}$$

where (i) comes from $\mu^\dagger \in \Pi_{\max}$; and (ii) is due to the fact that Ψ_η is a non-positive function.

On the other hand, due to that $\Pi_{\max} = (F+u)\cap\mathbb{R}^{XA}_{\geq 0}$, we have

$$\begin{aligned}
&\left\langle \mu^t, \hat{\ell}^t\right\rangle + \Phi_\eta\left(-\hat{L}^t\right) - \Phi_\eta\left(-\hat{L}^{t-1}\right) \\
\overset{(i)}{=}& \left\langle \mu^t, \hat{\ell}^t\right\rangle + \Phi_\eta\left(\nabla\Psi_\eta\left(\mu^t\right) + \boldsymbol{g}^t - \hat{\ell}^t\right) - \Phi_\eta\left(\nabla\Psi_\eta\left(\mu^t\right) + \boldsymbol{g}^t\right) \\
\overset{(ii)}{=}& \left\langle \mu^t, \hat{\ell}_t\right\rangle + \Phi_\eta\left(\nabla\Psi_\eta\left(\mu^t\right) - \hat{\ell}^t\right) - \Phi_\eta\left(\nabla\Psi_\eta\left(\mu^t\right)\right) \\
\overset{(iii)}{\leq}& \left\langle \mu^t, \hat{\ell}_t\right\rangle + \Psi_\eta^\star\left(\nabla\Psi_\eta\left(\mu^t\right) - \hat{\ell}^t\right) - \Psi_\eta^\star\left(\nabla\Psi\left(\mu^t\right)\right) \\
\overset{(iv)}{=}& D_{\Psi_\eta^\star}\left(\nabla\Psi_\eta\left(\mu^t\right) - \hat{\ell}^t, \nabla\Psi_\eta\left(\mu^t\right)\right),
\end{aligned}$$

where (i) follows from $\hat{L}^{t-1} + \nabla\Psi_\eta(\mu^t) + \boldsymbol{g}^t = 0$ for $\boldsymbol{g}^t \in F^\perp$; (ii) is due to the fact that $\boldsymbol{y} \in \mathbb{R}^{XA}$,

$$\begin{aligned}
\Phi_\eta\left(\boldsymbol{y} + \boldsymbol{g}^t\right) &= \sup_{\boldsymbol{\mu}\in(F+u)\cap\mathbb{R}^{XA}_{\geq 0}} \left\langle \boldsymbol{\mu}, \boldsymbol{y} + \boldsymbol{g}^t\right\rangle - \Psi_\eta\left(\boldsymbol{\mu}\right) \\
&= \left(\sup_{\boldsymbol{\mu}\in F\cap\mathbb{R}^{XA}_{\geq 0}} \left\langle \boldsymbol{\mu}, \boldsymbol{y} + \boldsymbol{g}^t\right\rangle - \Psi_\eta\left(\boldsymbol{\mu}\right)\right) + \left\langle u, \boldsymbol{y} + \boldsymbol{g}^t\right\rangle \\
&= \left(\sup_{\boldsymbol{\mu}\in F\cap\mathbb{R}^{XA}_{\geq 0}} \langle \boldsymbol{\mu}, \boldsymbol{y}\rangle - \Psi_\eta\left(\boldsymbol{\mu}\right)\right) + \left\langle u, \boldsymbol{y} + \boldsymbol{g}^t\right\rangle \\
&= \left(\sup_{\boldsymbol{\mu}\in(F+u)\cap\mathbb{R}^{XA}_{\geq 0}} \langle \boldsymbol{\mu}, \boldsymbol{y}\rangle - \Psi_\eta\left(\boldsymbol{\mu}\right)\right) + \left\langle u, \boldsymbol{g}^t\right\rangle \\
&= \Phi_\eta(\boldsymbol{y}) + \left\langle u, \boldsymbol{g}^t\right\rangle;
\end{aligned}$$

(iii) is by the observation that $\forall \boldsymbol{y} \in \mathbb{R}^{XA}, \Phi_\eta(\boldsymbol{y}) \leq \Psi_\eta^\star(\boldsymbol{y})$ and $\mu^t = \operatorname{argmax}_{\boldsymbol{x}\in\mathbb{R}^{XA}_{\geq 0}} \langle \boldsymbol{x}, \nabla\Psi_\eta(\mu^t)\rangle - \Psi_\eta(\boldsymbol{x})$ which implies that $\Phi_\eta(\nabla\Psi_\eta(\mu^t)) = \Psi_\eta^\star(\nabla\Psi_\eta(\mu^t))$; and (iv) comes from the definition of $D_{\Psi_\eta^\star}(\boldsymbol{x}, \boldsymbol{y})$. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. Combining Lemma 3, 4 and 7, with $p^\star$ computed in Algorithm 3, we have that

$$\begin{aligned}\mathfrak{R}_{\max}^T &\leq \textsc{Penalty} + \textsc{Stability} \\ &\leq \frac{H}{\eta}\log\left(XA\right) + \frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star \rho}\right)\lambda d H T\,, \end{aligned} \tag{9}$$

which along with choosing $\eta = \sqrt{\frac{2\log(AX)}{Td}}$ finishes the proof. □

B.1 BOUNDING THE PENALTY TERM

The lemma below upper bounds the PENALTY term.

Lemma 4. *For any fixed learning rate η and transition probability $p^\star$ over infoset-action space, it holds that*

$$\textsc{Penalty} \leq \frac{H}{\eta}\log\left(XA\right)\,.$$

Proof. It is clear that

$$-\Psi_\eta(\mu) = -\frac{1}{\eta}\sum_{h=1}^H \Psi_h\left(p^\star_{1:h}\cdot\mu_{1:h}\right) \overset{(i)}{\leq} \frac{1}{\eta}\sum_{h=1}^H \log\left(X_h A\right) \leq \frac{1}{\eta}\sum_{h=1}^H \log\left(XA\right) = \frac{H}{\eta}\log\left(XA\right)\,,$$

where (i) comes from Lemma 8. □

B.2 BOUNDING THE STABILITY TERM

Before bounding the stability term, we first introduce the following lemma, which bounds the variance of the loss estimate.

Lemma 5. *For any $h \in [H]$ and any $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$, it holds that $|\hat{\ell}_h^t(x_h, a_h)| \leq \frac{L^2}{\rho}$.*

Proof. First notice that for any ν^t and any $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$, we have

$$\begin{aligned}&\left\|\boldsymbol{\phi}^{\nu^t}(x_h, a_h)\right\|_2 \\ =&\left\|-\sum_{(s_h,b_h)\in x_h\times\mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h), b_h)\boldsymbol{\phi}(s_h, a_h, b_h)\right\|_2 \\ \leq&\sum_{(s_h,b_h)\in x_h\times\mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h), b_h)\left\|\boldsymbol{\phi}(s_h, a_h, b_h)\right\|_2 \\ \overset{(i)}{\leq}& L\sum_{(s_h,b_h)\in x_h\times\mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h), b_h) \\ \overset{(ii)}{\leq}& L\,,\end{aligned} \tag{10}$$

where (i) is due to Assumption 1; and (ii) follows from the proof of Lemma 2 by Kozuno et al. [2021].

Let $\Phi_h^t := \left\{\boldsymbol{\phi}^{\nu^t}(x_h, a_h)\right\}_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}$. It is then clear that

$$
\begin{aligned}
|\hat{\ell}_h^t(x_h,a_h)| &= |\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top (\boldsymbol{Q}_{\mu^t,h}^t)^{-1} \boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t) r_h^t(s_h^t,a_h^t,b_h^t)| \\
&\overset{(i)}{\le} |\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top (\boldsymbol{Q}_{\mu^t,h}^t)^{-1} \boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)| \\
&\overset{(ii)}{\le} \|\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\|_{(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}} \cdot \|\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)\|_{(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}} \\
&\le \|\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\|_{(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}} \cdot \sup_{\boldsymbol{\phi}\in\Phi_h^t} \|\boldsymbol{\phi}\|_{(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}} \\
&\le \sup_{\boldsymbol{\phi}\in\Phi_h^t} \|\boldsymbol{\phi}\|^2_{(\boldsymbol{Q}_{\mu^t,h}^t)^{-1}} \\
&\le \sup_{\boldsymbol{\phi}\in\Phi_h^t} \|\boldsymbol{\phi}\|^2_{(\rho\boldsymbol{I})^{-1}} \\
&\overset{(iii)}{\le} \frac{L^2}{\rho},
\end{aligned}
$$

where (i) is because $|r_h^t(s_h^t,a_h^t,b_h^t)| \le 1$; (ii) is by the Cauchy-Schwarz inequality; and (iii) comes from Eq. (10). □

Recall $\beta_h^\star = \min_{x_h\in\mathcal{X}_h} p_{1:h}^\star(x_h)$ and $\beta_h^\nu = \max_{t\in[T],x_h\in\mathcal{X}_h} p_{1:h}^{\nu^t}(x_h)$. The following lemma shows that the STABILITY term can be bounded by the variance of the loss estimate.

Lemma 6. *The one-step stability term satisfies*

$$
D_{\Psi_\eta^\star}(\nabla\Psi_\eta(\mu^t) - \hat{\ell}^t, \nabla\Psi_\eta(\mu^t)) \le \frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star \rho}\right) \sum_{h=1}^H \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}^t(x_h,a_h)}{p_{1:h}^\star(x_h)} \hat{\ell}_h^t(x_h,a_h)^2 .
$$

Proof. In what follows, we let $\tilde{\mu}^{t+1} := \nabla\Psi_\eta^\star\left(\nabla\Psi_\eta(\mu^t) - \hat{\ell}^t\right)$. Thus it is clear that

$$
\tilde{\mu}^{t+1} = \operatorname*{arg\,min}_{\mu\in\mathbb{R}_{\ge 0}^{XA}} \langle \mu, \hat{\ell}^t\rangle + D_{\Psi_\eta}(\mu,\mu^t),
$$

and

$$
\nabla\Psi_\eta(\tilde{\mu}^{t+1}) = \nabla\Psi_\eta(\mu^t) - \hat{\ell}^t, \tag{11}
$$

where the latter follows from the first-order optimality condition.

Then, one can deduce that

$$
\begin{aligned}
& D_{\Psi_\eta^\star}\left(\nabla\Psi_\eta(\mu^t) - \hat{\ell}^t, \nabla\Psi_\eta(\mu^t)\right) \\
&= D_{\Psi_\eta}\left(\nabla\Psi_\eta^\star\left(\nabla\Psi_\eta(\mu^t)\right), \nabla\Psi_\eta^\star\left(\nabla\Psi_\eta(\mu^t) - \hat{\ell}^t\right)\right) \\
&= D_{\Psi_\eta}\left(\mu^t, \tilde{\mu}^{t+1}\right) \\
&= -D_{\Psi_\eta}\left(\tilde{\mu}^{t+1}, \mu^t\right) + \left\langle \mu^t - \tilde{\mu}^{t+1}, \nabla\Psi_\eta(\mu^t) - \nabla\Psi_\eta(\tilde{\mu}^{t+1})\right\rangle \\
&\overset{(i)}{\le} -D_{\Psi_\eta}\left(\tilde{\mu}^{t+1}, \mu^t\right) + \frac{1}{2}\|\nabla\Psi_\eta(\mu^t) - \nabla\Psi_\eta(\tilde{\mu}^{t+1})\|_{\nabla^{-2}\Psi_\eta(z^t)} + \frac{1}{2}\|\mu^t - \tilde{\mu}^{t+1}\|_{\nabla^2\Psi_\eta(z^t)} \\
&\overset{(ii)}{=} -D_{\Psi_\eta}\left(\tilde{\mu}^{t+1}, \mu^t\right) + \frac{1}{2}\|\hat{\ell}^t\|_{\nabla^{-2}\Psi_\eta(z^t)} + \frac{1}{2}\|\mu^t - \tilde{\mu}^{t+1}\|_{\nabla^2\Psi_\eta(z^t)} \\
&\overset{(iii)}{=} \frac{1}{2}\|\hat{\ell}^t\|_{\nabla^{-2}\Psi_\eta(z^t)} \\
&\overset{(iv)}{=} \frac{\eta}{2}\sum_{h=1}^H \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{z_{1:h}^t(x_h,a_h)}{p_{1:h}^\star(x_h)} \hat{\ell}_h^t(x_h,a_h)^2 ,
\end{aligned} \tag{12}
$$

where (i) is by the Cauchy-Schwarz inequality and $z^t = \alpha\mu^t + (1-\alpha)\tilde{\mu}^{t+1}$ for some $\alpha \in [0,1]$; (ii) follows from Eq. (11); (iii) is due to the mean value theorem; and (iv) is by noticing that $\frac{\partial^2 \Psi_\eta(z_t)}{\partial z^t_{1:h}(x_h,a_h)^2} = \frac{p^\star_{1:h}(x_h)}{\eta z^t_{1:h}(x_h,a_h)}$.

Further, using Eq. (11) again and noticing that $\frac{\partial \Psi_\eta(mu_t)}{\partial \mu^t_{1:h}(x_h,a_h)} = \frac{p^\star_{1:h}(x_h)}{\eta}\left(\log p^\star_{1:h}(x_h)\mu^t_{1:h}(x_h,a_h) + 1\right)$, one can see that $\forall (x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}$,

$$\tilde{\mu}^{t+1}_{1:h}(x_h,a_h) = \mu^t_{1:h}(x_h,a_h)\exp\left(-\frac{\eta\hat{\ell}^t_h(x_h,a_h)}{p^\star_{1:h}(x_h)}\right).$$

The above display along with the fact that z^t in Eq. (12) satisfies $z^t = \alpha\mu^t + (1-\alpha)\tilde{\mu}^{t+1}$ for some $\alpha \in [0,1]$ implies that $\forall (x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}$,

$$z^t_{1:h}(x_h,a_h) \in \left[\mu^t_{1:h}(x_h,a_h)\exp\left(-\frac{\eta|\hat{\ell}^t_h(x_h,a_h)|}{p^\star_{1:h}(x_h)}\right), \mu^t_{1:h}(x_h,a_h)\exp\left(\frac{\eta|\hat{\ell}^t_h(x_h,a_h)|}{p^\star_{1:h}(x_h)}\right)\right]. \tag{13}$$

Combining Eq. (13) and Lemma 5 leads to

$$\mu^t_{1:h}(x_h,a_h)\exp\left(-\frac{\eta L^2}{p^\star_{1:h}(x_h)\rho}\right) \le z^t_{1:h}(x_h,a_h) \le \mu^t_{1:h}(x_h,a_h)\exp\left(\frac{\eta L^2}{p^\star_{1:h}(x_h)\rho}\right). \tag{14}$$

Substituting Eq. (14) into Eq. (12), we have

$$\begin{aligned}
D_{\Psi^\star_\eta}\left(\nabla\Psi_\eta\left(\mu^t\right) - \hat{\ell}^t, \nabla\Psi_\eta\left(\mu^t\right)\right) \le& \frac{\eta}{2}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{z^t_{1:h}(x_h,a_h)}{p^\star_{1:h}(x_h)}\hat{\ell}^t_h(x_h,a_h)^2 \\
\le& \frac{\eta}{2}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \exp\left(\frac{\eta L^2}{p^\star_{1:h}(x_h)\rho}\right)\frac{\mu^t_{1:h}(x_h,a_h)}{p^\star_{1:h}(x_h)}\hat{\ell}^t_h(x_h,a_h)^2 \\
\le& \frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta^\star_H\rho}\right)\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu^t_{1:h}(x_h,a_h)}{p^\star_{1:h}(x_h)}\hat{\ell}^t_h(x_h,a_h)^2,
\end{aligned}$$

which completes the proof. □

With "balanced transition" $p^\star$ computed in Algorithm 3, the following lemma upper bounds the STABILITY term via λ and d by considering the ratio between transition $p^{\nu^t}_{1:h}$ contributed by the environment state transition $\mathbb{P}$ as well as opponent's policy ν^t and "balanced transition" $p^\star$.

Lemma 7. *With "balanced transition" $p^\star$ computed in Algorithm 3, for any fixed learning rate η, it holds that*

$$\text{STABILITY} \le \frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta^\star_H\rho}\right)\lambda dHT.$$

Proof. Using Lemma 6, one can see that

$$
\begin{aligned}
&\textsc{Stability}\\
=&\mathbb{E}\left[\sum_{t=1}^{T} D_{\Psi_\eta^\star}(\nabla\Psi_\eta(\mu^t)-\hat{\ell}^t,\nabla\Psi_\eta(\mu^t))\right]\\
\leq&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\frac{1}{p^\star_{1:h}(x_h)}\mathbb{E}^{\mu^t,\nu^t}\left[\mu^t_{1:h}(x_h,a_h)\hat{\ell}^t(x_h,a_h)^2\right]\right]\\
=&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\frac{1}{p^\star_{1:h}(x_h)}\mu^t_{1:h}(x_h,a_h)\mathbb{E}^{\mu^t,\nu^t}\left[(\hat{\boldsymbol{\theta}}_h^t)^\top\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\hat{\boldsymbol{\theta}}_h^t\right]\right]\\
\overset{(i)}{=}&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\frac{1}{p^\star_{1:h}(x_h)}\mu^t_{1:h}(x_h,a_h)\right.\\
&\left.\cdot\mathbb{E}^{\mu^t,\nu^t}\left[r_h^t(s_h^t,a_h^t,b_h^t)^2\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)\right]\right]\\
\overset{(ii)}{\leq}&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\frac{1}{p^\star_{1:h}(x_h)}\mu^t_{1:h}(x_h,a_h)\right.\\
&\left.\cdot\mathbb{E}^{\mu^t,\nu^t}\left[\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)\right]\right]\\
\overset{(iii)}{\leq}&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{\beta_H^\star}\right.\\
&\left.\cdot\mathbb{E}^{\mu^t,\nu^t}\left[\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\right)(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)\right]\right]\\
=&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{\beta_H^\star}\mathbb{E}^{\mu^t,\nu^t}\left[\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\boldsymbol{\phi}^{\nu^t}(x_h^t,a_h^t)\right]\right]\\
=&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{\beta_H^\star}\operatorname{tr}\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}p^{\nu^t}_{1:h}(x_h)\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\right)\right]\\
\overset{(iv)}{\leq}&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\lambda\operatorname{tr}\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top(\boldsymbol{Q}^t_{\mu^t,h})^{-1}\right)\right]\\
=&\mathbb{E}\left[\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\lambda\operatorname{tr}(\boldsymbol{I}_d)\right]\\
=&\frac{\eta}{2}\exp\left(\frac{\eta L^2}{\beta_H^\star\rho}\right)THd\lambda\,,
\end{aligned}
$$

where (i) comes from the definition of $\hat{\boldsymbol{\theta}}_h^t$ in Eq. (4); (ii) holds due to that $|r_h^t(s_h^t,a_h^t,b_h^t)|\leq 1$; (iii) is because $\beta_H^\star\leq p^\star_{1:h}(x_h)$ for any $x_h\in\mathcal{X}_h$ and $h\in[H]$; and (iv) comes from the definition of λ.

The proof is thus concluded. □

B.3 PROPERTIES OF BALANCED TRANSITION $p^\star$

The lemma below delineates the property of $p^\star$ as transition probability over infoset-action space.

Lemma 8. *For any $h \in [H]$, any $p^\star$ as transition probability over infoset-actions and any policy $\mu \in \Pi_{\max}$ of the max-player, it holds that*

$$\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} p^\star_{1:h}(x_h)\mu_{1:h}(x_h,a_h) = 1\,.$$

Proof. By the definition of perfect recall and transition probability over infoset-action space, we have

$$\begin{aligned}
\mathbb{P}^{\mu,\nu}(x_h,a_h) &= \mathbb{P}^{\mu,\nu}(x_1,\ldots,x_h,a_h)\\
&= p^\star_0(x_1)\prod_{h'=1}^{h-1} p^\star_{h'}(x_{h'+1}|x_{h'},a_{h'})\cdot\prod_{h'=1}^{h}\mu_{h'}(a_{h'}|x_{h'})\\
&= p^\star(x_h)\mu_{1:h}(x_h,a_h)\,.
\end{aligned}$$

The proof is thus concluded by noticing that $\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mathbb{P}^{\mu,\nu}(x_h,a_h)=1$. □

The following lemma shows that $p^\star$ computed in Algorithm 3 guarantees $\lambda \le 1$ when the environment state transition is uniformly at random and the game tree is a k-ary tree.

Lemma 9. *When the environment state transition $\mathbb{P}$ is a uniform distribution and the game tree is a k-ary tree, the "balanced transition" $p^\star$ computed in Algorithm 3 guarantees that $\lambda \le 1$.*

Proof. Let $n, m > 0$ be the number of the states and the number of the infosets that the game can transit to when taking any action a. Note that $m \le n$ due to the properties of the perfect recall and tree structure conditions defined in Section 2.

Fix some $h \in [H]$. Recall $p^{\nu^t}_{1:h}(x_h) = \sum_{s_h:x(s_h)=x_h} p_{1:h}(s_h)\,\nu^t_{1:h-1}(y(s_{h-1}),b_{h-1})$. Since the environment state transition $\mathbb{P}$ is a uniform distribution, it holds that

$$\begin{aligned}
p^{\nu^t}_{1:h}(x_h) &= \sum_{s_h:x(s_h)=x_h} p_{1:h}(s_h)\,\nu^t_{1:h-1}(y(s_{h-1}),b_{h-1})\\
&= \left(\frac{1}{n}\right)^{h-1}\sum_{s_h:x(s_h)=x_h}\nu^t_{1:h-1}(y(s_{h-1}),b_{h-1})\\
&\le \left(\frac{1}{n}\right)^{h-1}\left(\frac{n}{m}\right)^{h-1}\\
&= \left(\frac{1}{m}\right)^{h-1}, \qquad (15)
\end{aligned}$$

where the inequality is by noticing that $\nu^t_{1:h-1}(y(s_{h-1}),b_{h-1}) \le 1$ for all $(y(s_{h-1}),b_{h-1}) \in \mathcal{Y}_h\times\mathcal{B}$. On the other hand, it is easy to check that $p^\star$ computed in Algorithm 3 satisfies $p^\star_{1:h}(x_h) = \left(\frac{1}{m}\right)^{h-1}$ for all $x_h \in \mathcal{X}_h$, which together with Eq. (15) concludes the proof. □

In the worst case, $p^{\nu^t}_{1:H}(x_H) = 1$ for some $x_H \in \mathcal{X}_H$ (note again that this is almost impossible to happen in practice as discussed in Section 4), meaning that $\lambda = \min_{x_H\in\mathcal{X}_H} 1/p^\star_{1:H}(x_H)$. Intuitively, λ can be well-controlled if the "balanced transition" $p^\star_{1:h}(\cdot)$ is "balanced" enough in the sense that the "transition probability" of visiting x_h is lower bounded for any $x_h \in \mathcal{X}_h$ and $h \in [H]$. This is exactly guaranteed by the design of our "balanced transition" $p^\star_{1:h}(\cdot)$ specified in Eq. (6), the computation of which is solved by our Algorithm 3. This is formalized by the following lemma.

Lemma 10. *The "balanced transition" $p^\star$ computed in Algorithm 3 guarantees that $\lambda \le X$.*

Proof. It suffices to show that $p^\star_{1:h}(x_h) \ge 1/X$ for any $x_h \in \mathcal{X}_h$ and $h \in [H]$. Clearly, $p^\star_{1:h}(\cdot)$ is minimzed at $h = H$ for some $x_H \in \mathcal{X}_H$ by its definition. By the construction of $p^\star_{1:h}(\cdot)$ in Algorithm 3, one can deduce that $\forall x_H \in \mathcal{X}_H$, we have

(understanding $\{(x_h, a_h)\}_{h\in[H-1]}$ as the unique trajectory leading to x_H below)

$$
\begin{aligned}
p^\star_{1:H}(x_H) &= p[x_H] \\
&= p\left[x_{H-1}\right] \cdot \frac{f\left[x_H\right]}{\sum_{x'_H \in C(x_{H-1}, a_{H-1})} f\left[x'_H\right]} \\
&= p\left[x_{H-2}\right] \cdot \frac{f\left[x_{H-1}\right]}{\sum_{x'_{H-1} \in C(x_{H-2}, a_{H-2})} f\left[x'_{H-1}\right]} \cdot \frac{f\left[x_H\right]}{\sum_{x'_H \in C(x_{H-1}, a_{H-1})} f\left[x'_H\right]} \\
&= p\left[x_{H-2}\right] \cdot \frac{f\left[x_{H-1}\right]}{\sum_{x'_{H-1} \in C(x_{H-2}, a_{H-2})} f\left[x'_{H-1}\right]} \cdot \frac{f\left[x_H\right]}{C\left[x_{H-1}, a_{H-1}\right]} \\
&\overset{(i)}{\geq} p\left[x_{H-2}\right] \cdot \frac{f\left[x_{H-1}\right]}{\sum_{x'_{H-1} \in C(x_{H-2}, a_{H-2})} f\left[x'_{H-1}\right]} \cdot \frac{f\left[x_H\right]}{f[x_{H-1}]} \\
&= p\left[x_{H-2}\right] \cdot \frac{f\left[x_H\right]}{\sum_{x'_{H-1} \in C(x_{H-2}, a_{H-2})} f\left[x'_{H-1}\right]} \\
&\geq \dots \\
&\geq \frac{f\left[x_H\right]}{\sum_{x_1 \in \mathcal{X}_1} f\left[x_1\right]} \\
&\geq \frac{f\left[x_H\right]}{X_H} \\
&\geq \frac{f\left[x_H\right]}{X} \\
&= \frac{1}{X},
\end{aligned}
$$

where (i) is due to $f[x_{H-1}] = \max_{a\in\mathcal{A}} C[x_{H-1}, a] \geq C[x_{H-1}, a_{H-1}]$ in Algorithm 3.

The proof is thus completed. □

C COMPUTATION ISSUE

In this section, we present the solutions to the optimization problem of the update of F^{2}TRL in Eq. (5) and the computation of the "balanced transition" $p^\star$ in Eq. (6).

C.1 F^{2}TRL UPDATE

To solve the update of F^{2}TRL in Eq. (5), we first present an OMD-like update as well as its solution. We then show that the solution to the OMD-like update is equivalent to the F^{2}TRL update, which provides the final optimization solution to the F^{2}TRL update.

To begin with, we first introduce the OMD-like update, which leverages a list of learning rates $\eta := (\eta_h(x_h))_{x_h\in\mathcal{X}_h, h\in[H]}$ adaptive to each infoset and a generalized potential function defined as follows (not to be confused with the negative entropy potential function $\Psi_h(\mu) = \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h, a_h) \log(\mu_{1:h}(x_h, a_h))$ used in F^{2}TRL):

$$
\psi_\eta(\mu) = \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}(x_h, a_h)}{\eta_h(x_h)} \log\left(\frac{\mu_{1:h}(x_h, a_h)}{\sum_{a'_h\in\mathcal{A}} \mu_{1:h}(x_h, a'_h)}\right).
$$

By the fact that for any $\mu \in \Pi_{\max}$, the derivative of $\psi_\eta(\mu)$ satisfies

$$
\nabla_{x_h,a_h} \psi_\eta(\mu) = \frac{1}{\eta_h(x_h)} \log(\mu_h(a_h|x_h)),
$$

one can see that $\psi_\eta(\mu)$ induces the following distance-generating function, which is a generalized version of the *dilated* entropy distance-generating function of Kozuno et al. [2021]:

$$D_{\psi_\eta}(\mu^1\|\mu^2) = \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu^1_{1:h}(x_h,a_h)}{\eta_h(x_h)} \log \frac{\mu^1_h(a_h|x_h)}{\mu^2_h(a_h|x_h)}\,. \tag{16}$$

The OMD-like update in accordance with the generalized dilated entropy distance-generating function in Eq. (16) is defined as

$$\begin{aligned}\mu^{t+1} &= \arg\min_{\mu\in\Pi_{\max}} \left\langle \mu, \hat{\ell}^t \right\rangle + D_{\psi_\eta}(\mu\|\mu^t) \\ &= \arg\min_{\mu\in\Pi_{\max}} \left\langle \mu, \hat{\ell}^t \right\rangle + \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}(x_h,a_h)}{\eta_h(x_h)} \log \frac{\mu_h(a_h|x_h)}{\mu^t_h(a_h|x_h)}\,.\end{aligned} \tag{17}$$

The solution to Eq. (17) is given in the following proposition. Notice that the solution to similar optimization problems of previous works [Kozuno et al., 2021, Bai et al., 2022, Fiegel et al., 2023] critically relies on the sparsity of their importance-weighted loss estimator, which only permits non-zero loss estimates along the experienced trajectory $\{(x^t_h, a^t_h)\}_{h\in[H]}$. In contrast, the solution to Eq. (17) in the following proposition solves the optimization problem of OMD with generalized dilated entropy distance-generating function and the loss estimator with non-zero loss estimates for all infoset-action pairs $(x,a)\in\mathcal{X}\times\mathcal{A}$.

Proposition 1. *The solution to the OMD-like update in Eq. (17) satisfies*

$$\begin{aligned}&\mu^{t+1}_h(a_h|x_h) \\ =&\mu^t_h(a_h|x_h)\exp\left\{-\eta_h(x_h)\hat{\ell}^t_h(x_h,a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h(x_h)}{\eta_{h+1}(x_{h+1})}\log Z^t_{h+1}(x_{h+1}) - \log Z^t_h(x_h)\right\},\end{aligned}$$

where

$$Z^t_h(x_h) = \sum_{a_h\in\mathcal{A}} \mu^t_h(a_h|x_h)\exp\left\{-\eta_h(x_h)\hat{\ell}^t_h(x_h,a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h(x_h)}{\eta_{h+1}(x_{h+1})}\log Z^t_{h+1}(x_{h+1})\right\}, \tag{18}$$

and for notational convenience, we define that $\forall(x_H,a_H)\in\mathcal{X}_H\times\mathcal{A}$, it has a unique descendant x_{H+1} such that $Z^t_{H+1}(x_{H+1})=1$.

Proof. First note that

$$\begin{aligned}&\left\langle \mu, \hat{\ell}^t \right\rangle + D_{\Psi_\eta}(\mu\|\mu^t) \\ =&\sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h,a_h)\left[\hat{\ell}^t_h(x_h,a_h) + \frac{1}{\eta_h(x_h)}\log\frac{\mu_h(a_h|x_h)}{\mu^t_h(a_h|x_h)}\right] \\ =&\sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_h} \mu_{1:h-1}(x_h)\left[\left\langle \mu_h(\cdot|x_h), \hat{\ell}^t_h(x_h,\cdot)\right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_h(\cdot|x_h)\|\mu^t_h(\cdot|x_h)\right)}{\eta_h(x_h)}\right].\end{aligned} \tag{19}$$

We now prove the proposition via backward induction over $h=H,\dots,1$.

When $h=H$, for any $x_H\in\mathcal{X}_H$, Eq. (19) shows that

$$\mu^{t+1}_H(a_H|x_H) = \mu^t_H(a_H|x_H)\exp\left\{-\eta_h(x_h)\hat{\ell}^t_H(x_H,a_H) - \log Z^t_H(x_H)\right\},$$

where $Z^t_H(x_H) = \sum_{a_H\in\mathcal{A}}\mu^t_H(a_H|x_H)\exp\{-\eta_h(x_H)\hat{\ell}^t_H(x_H,a_H)\}$ is a normalization factor.

Fix some $h \in [H]$. Now suppose the induction hypothesis holds from step $h+1$ to H and consider the h-th step. Using the induction hypothesis, one can see that Eq. (19) can be rewritten as

$$
\begin{aligned}
&\sum_{h'=1}^{H} \sum_{(x_{h'},a_{h'})\in\mathcal{X}_{h'}\times\mathcal{A}} \mu_{1:h'}(x_{h'},a_{h'}) \left[\hat{\ell}_{h'}^t\left(x_{h'},a_{h'}\right) + \frac{1}{\eta_{h'}(x_{h'})} \log \frac{\mu_{h'}(a_{h'}|x_{h'})}{\mu_{h'}^t(a_{h'}|x_{h'})}\right] \\
&= \sum_{h'=1}^{H} \sum_{x_{h'}\in\mathcal{X}_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \hat{\ell}_{h'}^t\left(x_{h'},\cdot\right)\right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_{h'}(\cdot|x_{h'})\|\mu_{h'}^t(\cdot|x_{h'})\right)}{\eta_{h'}(x_{h'})}\right] \\
&= \sum_{h'=1}^{h} \sum_{x_{h'}\in\mathcal{X}_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \hat{\ell}_{h'}^t\left(x_{h'},\cdot\right)\right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_{h'}(\cdot|x_{h'})\|\mu_{h'}^t(\cdot|x_{h'})\right)}{\eta_{h'}(x_{h'})}\right] \\
&\quad + \sum_{h'=h+1}^{H} \left[\sum_{x_{h'+1}\in\mathcal{X}_{h'+1}} \frac{\mu_{1:h'}(x_{h'+1})}{\eta_{h'+1}(x_{h'+1})} \log Z_{h'+1}^t(x_{h'+1}) - \sum_{x_{h'}\in\mathcal{X}_{h'}} \frac{\mu_{1:h'-1}(x_{h'})}{\eta_{h'}(x_{h'})} \log Z_{h'}^t(x_{h'})\right] \\
&= \sum_{h'=1}^{h} \sum_{x_{h'}\in\mathcal{X}_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \hat{\ell}_{h'}^t\left(x_{h'},\cdot\right)\right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_{h'}(\cdot|x_{h'})\|\mu_{h'}^t(\cdot|x_{h'})\right)}{\eta_{h'}(x_{h'})}\right] \\
&\quad - \sum_{x_{h+1}\in\mathcal{X}_{h+1}} \frac{\mu_{1:h}(x_{h+1})}{\eta_{h+1}(x_{h+1})} \log Z_{h+1}^t(x_{h+1}) \\
&= \sum_{h'=1}^{h-1} \sum_{x_{h'}\in\mathcal{X}_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \hat{\ell}_{h'}^t\left(x_{h'},\cdot\right)\right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_{h'}(\cdot|x_{h'})\|\mu_{h'}^t(\cdot|x_{h'})\right)}{\eta_{h'}(x_{h'})}\right] \\
&\quad + \sum_{x_h\in\mathcal{X}_h} \mu_{1:h-1}(x_h) \Bigg[\underbrace{\left\langle \mu_h(\cdot|x_h), \hat{\ell}_h^t\left(x_h,\cdot\right) - \sum_{x_{h+1}\in C(x_h,\cdot)} \frac{\log Z_{h+1}^t(x_{h+1})}{\eta_{h+1}(x_{h+1})}\right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_h(\cdot|x_h)\|\mu_h^t(\cdot|x_h)\right)}{\eta_h(x_h)}}_{\heartsuit}\Bigg] .
\end{aligned}
$$

By minimizing ($\heartsuit$), one can derive that

$$
\begin{aligned}
&\mu_h^{t+1}(a_h|x_h) \\
=&\mu_h^t(a_h|x_h) \exp\left\{-\eta_h(x_h)\hat{\ell}_h^t(x_h,a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h(x_h)}{\eta_{h+1}(x_{h+1})} \log Z_{h+1}^t(x_{h+1}) - \log Z_h^t(x_h)\right\} ,
\end{aligned}
$$

where

$$
Z_h^t(x_h) = \sum_{a_h\in\mathcal{A}} \mu_h^t(a_h|x_h) \exp\left\{-\eta_h(x_h)\hat{\ell}_h^t(x_h,a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h(x_h)}{\eta_{h+1}(x_{h+1})} \log Z_{h+1}^t(x_{h+1})\right\} .
$$

The proof is thus concluded. □

In what follows, for notational convenience, we denote $J_h^t(x_h,a_h) = -\eta_h(x_h)\hat{\ell}_h^t(x_h,a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h(x_h)}{\eta_{h+1}(x_{h+1})} \log Z_{h+1}^t(x_{h+1})$ as the surrogate loss.

To solve the update of F^{2}TRL, we follow the same idea as Fiegel et al. [2023] that translates the update of FTRL into an OMD-like update. In specific, Proposition F.2 of Fiegel et al. [2023] shows that the update of Eq. (5) is equivalent to the solution to the following optimization problem:

$$
\mu^{t+1} = \operatorname*{arg\,min}_{\mu\in\Pi_{\max}} \left\langle \mu, \hat{L}^t\right\rangle + D_{\Psi_{\eta^\star}}\left(\mu\|\mu^\star\right) ,
$$

where $\eta^\star$ and $\mu^\star$ satisfy

$$
\eta_h^\star(x_h) = \frac{\eta}{(H-h+1)p_{1:h}^\star(x_h)} , \tag{20}
$$

Algorithm 2 F^{2}TRL Update

Input: Tree-like structure of $\mathcal{X} \times \mathcal{A}$, fixed learning rates η, "balanced transition" $p^\star$ and cumulative loss estimates $\left\{\hat{L}_h^t(x_h, a_h)\right\}_{(x_h,a_h)\in\mathcal{X}\times\mathcal{A}}$.
Initialization: For all x_H in $\mathcal{X}_H$, initialize $Z^t(x_{H+1}) = 1$. Set adaptive learning rates $\eta^\star$ according to Eq. (20). Set base policy $\mu^\star$ according to Eq. (21).
for $h = H$ to 1 **do**
  for x_h in $\mathcal{X}_h$ **do**
    Compute $J_h^t(x_h, a_h) = -\eta_h^\star(x_h)\hat{L}^t(x_h, a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h^\star(x_h)}{\eta_{h+1}^\star(x_{h+1})} \log Z_{h+1}^t(x_{h+1})$.
    Compute $Z_h^t(x_h) = \sum_{a_h\in\mathcal{A}} \mu_h^\star(a_h|x_h) \exp\left(J_h^t(x_h, a_h)\right)$.
    for a_h in $\mathcal{A}$ **do**
      Compute $\mu_h^{t+1}(a_h|x_h) = \mu_h^\star(a_h|x_h) \exp\left(J_h^t(x_h, a_h) - \log Z_h^t(x_h)\right)$.
    end for
  end for
end for

and

$$\mu^\star = \arg\min_{\mu\in\Pi_{\max}} \sum_{h=1}^{H} \Psi_h\left(p_{1:h}^\star \cdot \mu_{1:h}\right) , \tag{21}$$

and recall $D_{\psi_{\eta^\star}}\left(\mu^1, \mu^0\right) = \sum_{h=1}^H \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}^1(x_h,a_h)}{\eta_h^\star(x_h)} \log \frac{\mu_h^1(a_h|x_h)}{\mu_h^0(a_h|x_h)}$. Note that $\mu^\star$ can be computed efficiently via backward dynamic programming.

Then, combined with the solution to the OMD-like update in Proposition 1, the solution to the update of F^{2}TRL can be obtained by substituting $\hat{\ell}^t$ and μ^t with $\hat{L}^t$ and $\mu^\star$ in Eq. (17), the details of which are presented in Algorithm 2.

Algorithm 3 Compute $p^\star$

Input: Tree-like structure of $\mathcal{X} \times \mathcal{A}$.
Initialization: Transition array $p[\cdot]$ of size X ; auxiliary array $f[\cdot]$ of size X, $C[\cdot,\cdot]$ of size $X \times A$. For all x_H in $\mathcal{X}_H$, initialize $f[x_H] = 1$.
for $h = H-1$ to 1 **do**
  for x_h in $\mathcal{X}_h$ **do**
    for a_h in $\mathcal{A}$ **do**
      Compute $C[x_h, a_h] = \sum_{x_{h+1}\in C(x_h,a_h)} f[x_{h+1}]$,
    end for
    Compute $f[x_h] = \max_{a\in\mathcal{A}} C[x_h, a]$.
  end for
end for
for x_1 in $\mathcal{X}_1$ **do**
  Compute $p[x_1] = \frac{f[x_1]}{\sum_{x_1\in\mathcal{X}_1} f[x_1]}$.
end for
for $h = 1$ to $H-1$ **do**
  for x_h, a_h in $\mathcal{X}_h \times \mathcal{A}$ **do**
    for x_{h+1} in $C(x_h, a_h)$ **do**
      Compute $p[x_{h+1}] = p[x_h] \cdot \frac{f[x_{h+1}]}{\sum_{x_{h+1}\in C(x_h,a_h)} f[x_{h+1}]}$.
    end for
  end for
end for
return p.

C.2 COMPUTATION OF BALANCED TRANSITION $p^\star$

This section presents Algorithm 3, which solves the computation of $p^\star$ defined in Eq. (6) via backward dynamic programming.

D PROOF OF REGRET LOWER BOUND

In this section, we present the proof of Theorem 2.

Proof of Theorem 2. We consider an A-ary tree IIEFG instance, in which

- $B = 1$ so that there is actually no opponent effectively (and hence the dependence on the opponent's action b is omitted in what follows);
- $X_h = S_h = A^{h-1}$ for all $h \in [H]$, which means that $\mathcal{X}_h = \mathcal{S}_h$ and there is actually no partial observability;
- $r_h(s, a) = 0$ for all $h \in [H-1]$, and $r_H(s, a)$ is a reward sampled from Bernoulli distribution $\mathrm{Ber}(\bar{r}_H(s, a))$ with mean $\bar{r}_H(s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta} \rangle$.

By the construction, there exists a unique action sequence $(a_1, \ldots, a_{h-1})$ that determines s_h (and hence x_h) and the transition is deterministic and known. Following similar arguments by Bai et al. [2022], Fiegel et al. [2023], it can be shown that if algorithm Alg achieves regret $\mathfrak{R}^T_{\max}$ on this IIEFG instance, then Alg can be used to tackle a stochastic linear bandit problem with A^H "arms" and obtain the regret with the same order as $\mathfrak{R}^T_{\max}$, where the reward for "arm" $(a_1, a_2, \ldots, a_H)$ (*i.e.*, (s_H, a_H)) is sampled from $\mathrm{Ber}(\langle \boldsymbol{\phi}(s_H, a_H), \boldsymbol{\theta} \rangle)$.

We now first consider the case when $H \geq d$. In this case, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ satisfy $\boldsymbol{\phi}(s, a)_{[1:d-1]} \in \{-1, 1\}^{d-1}$, $\boldsymbol{\phi}(s, a)_d = 1/4$, $\boldsymbol{\theta}_{[1:d-1]} \in \{-\Delta, \Delta\}^{d-1}$ with $\Delta = 1/(8\sqrt{2T})$ and $\boldsymbol{\theta}_d = 1$. Moreover, since $|\mathcal{S}_H \times \mathcal{A}| = A^{H-1} \cdot A = A^H$ as well as $H \geq d$ and $A \geq 2$, $\boldsymbol{\phi}$ can be chosen such that $\{\boldsymbol{\phi}(s, a)_{[1:d-1]}\}_{(s,a) \in \mathcal{S}_H \times \mathcal{A}} = \{-1, 1\}^{d-1}$ (omitting the duplicate feature vectors). Then by canonical analysis for the regret lower bound of stochastic linear bandits (see, *e.g.*, Theorem 24.1 by Lattimore and Szepesvári [2020]; Lemma 25 by Zhou et al. [2021]), there exists a $\boldsymbol{\theta}^{\mathrm{Alg}}_{[1:d-1]} \in \{-\Delta, \Delta\}^{d-1}$ such that $\mathfrak{R}^T \geq (d-1)\sqrt{T}/(16\sqrt{2}) = \Omega(\sqrt{d^2 T})$.

In case when $H < d$, we can choose $\boldsymbol{\phi}$ such that the stochastic linear bandit problem, on which Alg suffers the same regret as on the IIEFG instance, has 2^H distinct feature vectors since $A \geq 2$ and $A^H \geq 2^H$. Then by similar reasoning of the construction of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ in the case $H \geq d$ and the proof of Corollary 3 by Zhou [2019], there exists a $\boldsymbol{\theta}^{\mathrm{Alg}}$ such that $\mathfrak{R}^T \geq \Omega(\sqrt{dHT})$.

The proof is concluded by combining the results of the two cases.

□