TOWARDS PROVABLY EFFICIENT LEARNING OF EXTENSIVE-FORM GAMES WITH IMPERFECT INFORMATION AND LINEAR FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study two-player zero-sum imperfect information extensive-form games (IIEFGs) with linear functional approximation. In particular, we consider linear IIEFGs in the formulation of partially observable Markov games (POMGs) with unknown transition and bandit feedback, in which the reward function admits a linear structure. To tackle the partial observation of this problem, we propose a linear loss estimator based on the *composite* features of information set-action pairs. Through integrating this loss estimator with the online mirror descent (OMD) framework and delicate analysis of the stability term in the linear case, we prove the $\widetilde{\mathcal{O}}(\sqrt{(d+1/\rho)HX^2T})$ regret upper bound of our algorithm, where H is the horizon length, X is the cardinality of the information set space, d is the ambient dimension of the feature mapping, and ρ is the minimum eigenvalue of the feature covariance matrix generated by the exploration policy. Additionally, by leveraging the "transitions" over information set-actions, we propose another algorithm based on the follow-the-regularized-leader (FTRL) framework, attaining a regret bound of $\widetilde{\mathcal{O}}(\sqrt{H^2d\lambda T})$, where λ is a quantity depends on the game tree structure. Moreover, we prove that our FTRL-based algorithm also achieves the $\widetilde{\mathcal{O}}(\sqrt{HXdT})$ regret with a different initialization of parameters. Further, we provide an $\Omega(\sqrt{d\min(d,H)T})$ regret lower bound for this problem. To the best of our knowledge, we present the first line of algorithms studying learning IIEFGs with linear function approximation.

1 INTRODUCTION

In imperfect information games (IIGs), players are limited in their knowledge of the true state of play when making moves. This opacity allows for intricate strategic maneuvers like bluffing, as players can hide private information from opponents. In particular, the notion of imperfect-information extensive-form games (IIEFGs) (Kuhn, 1953) simultaneously enables imperfect information and sequential play, which thus characterizes a large amount of modeling real-world imperfect information games including Poker (Heinrich et al., 2015; Moravčík et al., 2017; Brown & Sandholm, 2018), Bridge (Tian et al., 2020), Scotland Yard (Schmid et al., 2021) and Mahjong (Li et al., 2020; Kurita & Hoki, 2021; Fu et al., 2022). There has been an extensive line of works on regret minimization or finding the Nash equilibrium (NE) (Nash Jr, 1950) of IIEFGs. Under perfect recall condition, when the full knowledge of the game is known, existing works solve this problem by linear programming (Koller & Megiddo, 1992; Von Stengel, 1996; Koller et al., 1996), first-order optimization methods (Hoda et al., 2010a; Kroer et al., 2015a; 2018; Munos et al., 2020; Lee et al., 2021; Liu et al., 2022), and counterfactual regret minimization (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 or only partial knowledge of the game is revealed, the problem will be much more challenging and is typically tackled through learning from the random observations accrued during repeated plays of the game. In this line of works, two-player zero-sum IIEFGs have been addressed via equipping online mirror descent (OMD) or follow-the-regularized-leader (FTRL) frameworks with loss estimations (Farina et al., 2021; Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023) and Monte-Carlo counterfactual regret minimization (Lanctot et al., 2009; Farina et al., 2020; Farina & Sandholm, 2021). Amongst these work, Bai

et al. (2022) leverage OMD with "balanced exploration policies" to achieve the first $\widetilde{\mathcal{O}}(\sqrt{H^3XAT})$ 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 sample complexity matches the information-theoretic lower bound on all factors but H up to logarithmic factors. Subsequently, Fiegel et al. (2023) further improve the bound to $\widetilde{\mathcal{O}}(\sqrt{XAT})$, with optimal dependency on all factors up to logarithmic factors, using FTRL with "balanced transitions".

Though significant advances have emerged in learning two-player zero-sum IIEFGs, the existing sample complexities of all works depend on X and A. In practice, however, X and/or A might be very large, particularly in large-scale IIEFGs, which makes the above sample complexities vacuous. This issue, which is typically called the *curse of dimensionality*, has also emerged in various problems beyond IIEFGs. To cope with this issue, a common approach is *function approximation*, which approximates the observations on experienced information sets and/or actions with sharing parameters and generalizes them onto unseen information sets and/or 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 algorithms with function approximation for IIEFGs still remain open and we are still far from understanding them well. On the other hand, in the more amenable sequential-decision making problems including (single-agent) reinforcement learning (RL) (Ayoub et al., 2020; Jin et al., 2020; Zhou et al., 2021; He et al., 2023) and Markov games (MGs) with perfect information (Chen et al., 2022; Ni et al., 2023; Wang et al., 2023; Cui et al., 2023), significant advances have emerged in understanding the theoretical guarantees of algorithms with function approximation. Therefore, the above two facts naturally motivate us to ask the following question:

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

In this paper, we give an affirmative answer to the above question for IIEFGs with linear function approximation, in the *offline* setting[1]. In specific, we consider IIEFGs in the formulation of partially observable Markov games (POMGs) with unknown transition and unknown rewards while admitting a linear structure over the reward functions. This problem is challenging in that both players are unaware of the current underlying state since only the current information set rather than the state is observable, which poses substantial difficulties in exploiting the linear structure of the reward functions, as the current feature corresponding to the current state is unknown. To address this problem and also establish efficient algorithms for learning IIEFGs with linear function approximation, in this paper, we make the following contributions:

- To learn the unknown parameter that linearly parameterizes the reward functions, we instead utilize a kind of *composite* reward features, weighted by the transitions and opponent's policy. Intuitively, composite reward features can be seen as features of corresponding information set-actions. Equipped with the composite reward features, we further propose the first least-squares loss estimator for this problem and prove its unbiasedness (see Section 3.1 for details).
- Based on the least-squares loss estimator, we then propose the least-squares online mirror descent (LSOMD) algorithm that attains the $\widetilde{\mathcal{O}}(\sqrt{(d+1/\rho)HX^2T})$ regret bound, where d is the ambient dimension of the feature mapping and $\rho := \min_{t\in[T],h\in[H]} \lambda_{\min}(\boldsymbol{Q}_{\pi^t,h})$ with $\boldsymbol{Q}_{\pi^t,h}$ being as the feature covariance matrix induced by exploration policy π^t at step h. Compared to the computation and regret analysis of OMD in tabular IIEFGs (Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023) that heavily depends on the sparsity of the importance-weighted loss estimate, however, our case intrinsically requires new ingredients to solve both aspects, due to the leverage of the linear structure. The key insight is to solve the computation and also bound the stability term of LSOMD by the log-partition function $\log Z_1^t$, which is in turn bounded by the expectation of the element-wise product of all the random vectors sampled from all the categorical distributions along paths from the root node (see Section 3.3 for details).
- Via integrating our proposed linear loss estimator, the solution to the optimization problem based on the log-partition function $\log Z_1^t$ and the idea of "balanced transition", which shares a similar spirit as Bai et al. (2022); Fiegel et al. (2023), we additionally propose the least-squares follow-the-regularized-leader (LSFTRL) algorithm. Let $p^{\nu}_{1:h}(x_h) =$

[1]By "offline" we refer to that the feature vectors of state-action weighted by min-player's policy ν^t in episode t (as well as transitions) are accessible to the max-player before the t-th episode starts. Please see Section 2 for more discussions.

Table 1: Comparisons of regret bounds with most related works studying IIEFGs when the full knowledge of the game is not known a priori.

Algorithm	Setting	Regret
`IXOMD` (Kozuno et al., 2021)	Online	$\widetilde{\mathcal{O}}(HX\sqrt{AT})$
`Balanced OMD/CFR` (Bai et al., 2022)		$\widetilde{\mathcal{O}}(\sqrt{H^3XAT})$
`Balanced FTRL` (Fiegel et al., 2023)		$\widetilde{\mathcal{O}}(\sqrt{XAT})$
`LSOMD` (this paper)	Offline[1]	$\widetilde{\mathcal{O}}(\sqrt{(d+1/\rho)HX^2T})$ [2]
`LSFTRL` (this paper)		$\widetilde{\mathcal{O}}(\sqrt{H^2d\lambda T})/\widetilde{\mathcal{O}}(\sqrt{HXdT})$ [3]
Lower bound (this paper)	-	$\Omega(\sqrt{d\min(d,H)T})$

[1] See Section 2 for the definition of our *offline* setting.
[2] See Assumption 3.2 for the definition of ρ.
[3] The λ in the former bound depends on the game tree structure, defined in Assumption 4.1. The latter bound is obtained by the same algorithm but with a different initiation of parameters.

$\sum_{s_h \in x_h} p_{1:h}(s_h)\nu_{1:h-1}(y(s_{h-1}), b_{h-1})$ be the sequence-form representation of the "transition" over $\mathcal{X}_h \times \mathcal{A} \times \mathcal{X}_{h+1}$ induced by the environment transition $\mathbb{P} = \{p_h\}_{h=0}^{H-1}$ and opponent's policy ν. Under the assumption that for any $t \in [T]$, $h \in [H]$ and $x_1, x_2 \in \mathcal{X}_h$, $p^{\nu^t}_{1:h}(x_1)/p^{\star}_{1:h}(x_2) \le \lambda$ with $p^{\star}_{1:h}$ being the sequence-form representation of the chosen "balanced transition" over $\mathcal{X}_h \times \mathcal{A} \times \mathcal{X}_{h+1}$ in `LSFTRL`, we prove that the regret upper bound of `LSFTRL` is of order $\widetilde{\mathcal{O}}(\sqrt{H^2d\lambda T})$. With a different initialization of "balanced transition" parameter that is not leveraged in previous works and a refined analysis on the stability term, we also prove that `LSFTRL` enjoys a $\widetilde{\mathcal{O}}(\sqrt{HXdT})$ regret (see Section 4.2 for details).

1.1 Related Work

Partially observable Markov games (POMGs) With perfect information, learning MGs dates back to the work of Littman & Szepesvári (1996) and has been well-studied (Littman, 2001; Greenwald & Hall, 2003; Hu & Wellman, 2003; Hansen et al., 2013; Sidford et al., 2018; Lagoudakis & Parr, 2002; Pérolat et al., 2015; Fan et al., 2020; Jia et al., 2019; Cui & Yang, 2021; Zhang et al., 2021; Bai & Jin, 2020; Liu et al., 2021; Zhou et al., 2021; Song et al., 2022; Li et al., 2022; Xiong et al., 2022; Wang et al., 2023; Cui et al., 2023). When only with imperfect information but the full model of the game (*i.e.*, transitions and rewards) is known, existing works can be categorized into three lines. The first line uses sequence-form policies to reformulate this problem as a linear program (Koller & Megiddo, 1992; Von Stengel, 1996; Koller et al., 1996). The second line considers solving the minimax optimization problem directly by first-order algorithms (Hoda et al., 2010a; Kroer et al., 2015a; 2018; Munos et al., 2020; Lee et al., 2021; Liu et al., 2022). The last line of works tackles this problem using counterfactual regret minimization (CFR), which minimizes counterfactual regrets locally at each information set (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 model of the game is not known or only partial knowledge of the game is accessible, the Monte-Carlo CFR algorithm proposed by Lanctot et al. (2009) attains the first ε-NE result in this problem. Subsequently, this framework is further generalized by Farina et al. (2020); Farina & Sandholm (2021). Besides, the other line of works considers combining OMD and FTRL with importance-weighted loss estimator (Farina et al., 2021; Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023) to tackle this problem. Remarkably, Bai et al. (2022) obtain the $\widetilde{\mathcal{O}}(\sqrt{H^3XAT})$ regret by using "balanced" dilated KL as the distance metric. With an analogous notion of "balanced" transition, Fiegel et al. (2023) finally achieve the regret of order $\widetilde{\mathcal{O}}(\sqrt{XAT})$, matching the lower bound up to logarithmic factors.

Markov games with Function Approximation To cope with the issue of the curse of dimensionality in MGs, there has been growing research interest in learning MGs in the function approximation setting recently (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). In particular, Xie et al. (2020) assume

both the transition and the reward functions of the episodic two-player zero-sum MGs are linearly realizable and achieve an $\widetilde{\mathcal{O}}(\sqrt{d^3H^4T})$ regret. More recent works generally fall into two categories. The first category aims to relax the assumption of linear function approximation by studying MGs in 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 *perfect information* MGs with function approximation, and (to our knowledge) there are no existing works studying *partially observable* MGs with function approximation, which is the main focus of our work.

2 PRELIMINARIES

Following previous works (Kozuno et al., 2021; Bai et al., 2022), in this work, we also study IIEFGs in the formulation of POMGs, the preliminaries of which are introduced in this section.

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

- H is the length of the horizon;
- $\mathcal{S} = \bigcup_{h\in[H]} \mathcal{S}_h$ is a finite state space with cardinality $S = \sum_{h=1}^{H} S_h$ and $|\mathcal{S}_h| = S_h$;
- $\mathcal{X} = \bigcup_{h\in[H]} \mathcal{X}_h$ and $\mathcal{Y} = \bigcup_{h\in[H]} \mathcal{Y}_h$ are the spaces of information sets (short for *infosets* in the following paper) for the *max-player* and *min-player*, respectively. Specifically, the cardinality X of $\mathcal{X}$ satisfies $X \coloneqq \sum_{h=1}^{H} X_h$ with $|\mathcal{X}_h| = X_h$ and the cardinality Y of $\mathcal{Y}$ satisfies $Y \coloneqq \sum_{h=1}^{H} Y_h$ with $|\mathcal{Y}_h| = Y_h$;
- $\mathcal{A}$ with $|\mathcal{A}| = A$ and $\mathcal{B}$ with $|\mathcal{B}| = B$ are the finite action spaces for the max-player and min-player, respectively;
- $\mathbb{P} = \{p_0(\cdot) \in \Delta_{\mathcal{S}_1}\} \cup \left\{p_h(\cdot \mid s_h, a_h, b_h) \in \Delta_{\mathcal{S}_{h+1}}\right\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}, h\in[H-1]}$ are the transition probability functions[2], with $p_0(\cdot)$ being the probability distribution of the initial states, and $p_h(s_{h+1}|s_h, a_h, b_h)$ being the probability of transmitting to the next state s_{h+1} conditioned on (s_h, a_h, b_h) at step h;
- $r = \{r_h(s_h, a_h, b_h) \in [-1, 1]\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}}$ are the stochastic reward functions with $\bar{r}_h(s_h, a_h, b_h)$ as means.

Learning Protocol To begin with, we denote by $\mu \coloneqq \{\mu_h\}_{h\in[H]}$ with $\mu_h^t : \mathcal{X}_h \to \Delta_{\mathcal{A}}$ the max-player's stochastic policy and by $\Pi_{\max}$ the set of the policies of the max-player. The min-player's stochastic policy ν and the set of the policies of the min-player $\Pi_{\min}$ are defined similarly. The game proceeds in T episodes. At the beginning of episode t, the max-player chooses a stochastic policy $\mu_t \in \Pi_{\max}$. And similarly, the min-player chooses $\nu_t \in \Pi_{\min}$. Then, an initial state s_1^t will be sampled from p_0. At each step h, the max-player, and min-player will observe their infoset $x_h^t \coloneqq x(s_h^t)$ and $y_h^t \coloneqq y(s_h^t)$ respectively, but *without* observing s_h^t. Conditioned on x_h^t, the max-player will sample and execute an action $a_h^t \sim \mu_h^t(\cdot|x_h)$. Simultaneously, the min-player will take action $b_h^t \sim \nu_h^t(\cdot|y_h)$. Subsequently, the game will transit to the next state s_{h+1}^t, which is drawn from $p_h(\cdot|s_h^t, a_h^t, b_h^t)$. Also, the max-player and min-player will receive rewards $r_h^t \coloneqq r_h(s_h^t, a_h^t, b_h^t)$ and $-r_h^t$ respectively. The episode will terminate after taking actions a_H^t and b_H^t conditioned on x_H^t and y_H^t respectively, *i.e.*, the game will terminate in H steps.

Perfect Recall and Tree Structure As in previous works (Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023), we also suppose that the POMGs satisfy the *tree structure* and the *perfect recall* condition (Kuhn, 1953). In specific, the tree structure indicates that for any $h = 2, \ldots, H$ and $s_h \in \mathcal{S}$, 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. Besides, perfect recall condition holds for each player if for any $h = 2, \ldots, H$ and any infoset $x_h \in \mathcal{X}_h$ of the max-player, 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 slightly abuse of notations, we also let $C_{h'}(x_h) \coloneqq \cup_{a_h\in\mathcal{A}} C_{h'}(x_h, a_h)$ and $C(x_h, a_h) \coloneqq C_{h+1}(x_h, a_h)$.

[2]While in some games, $\{p_h\}_{h=1}^{H-1}$ might be time-homogeneous, *i.e.*, $\{p_h\}_{h=1}^{H-1}$ does not depend on h, we retain the dependence on h in our notations as it allows the results to be applicable more broadly without too much additional efforts in the analysis, following previous works (Bai et al., 2022; Fiegel et al., 2023).

Sequence-form Representations In addition, 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):

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

where $p_{1:h}(s_h)$ is the sequence-form transition probability defined as $p_{1:h}(s_h) = p_0(s_1)\prod_{h'\le h-1} p_{h'}(s_{h'+1} \mid s_{h'}, a_{h'}, b_{h'})$, and $\mu_{1:h}(\cdot,\cdot)$ and $\nu_{1:h}(\cdot,\cdot)$ are the sequence-form policies satisfying $\mu_{1:h}(x_h, a_h) \coloneqq \prod_{h'=1}^{h} \mu_{h'}(a_{h'} \mid x_{h'})$ and $\nu_{1:h}(y_h, b_h) \coloneqq \prod_{h'=1}^{h} \nu_{h'}(b_{h'} \mid y_{h'})$. Therefore, we slightly abuse the meanings of μ and ν by viewing $\mu = \{\mu_{1:h}\}_{h\in[H]}$ and $\nu = \{\nu_{1:h}\}_{h\in[H]}$ as *realization plans* (Von Stengel, 1996). Under sequence-form representations, it is then clear that $\Pi_{\max}$ is a convex compact subspace of $\mathbb{R}^{XA}$ satisfying constraints $\mu_{1:h}(x_h, a_h) \ge 0$ 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 $\mu_{1:0}(x_0, a_0) = p(\emptyset) = 1$).

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

Assumption 2.1 (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 for any $(s_h, a_h, b_h) \in \mathcal{S}_h\times\mathcal{A}\times\mathcal{B}$, it holds that $\bar{r}_h(s_h, a_h, b_h) = \langle \boldsymbol{\phi}(s_h, a_h, b_h), \boldsymbol{\theta}_h\rangle$. In addition, we further assume that $\|\boldsymbol{\theta}_h\|_2 \le \sqrt{d}$, $\sup_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}} \|\boldsymbol{\phi}(s_h, a_h, b_h)\|_2 \le 1$, 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 $\mathbb{R}^d$, for any $h \in [H]$.*

Similar assumptions imposed over reward functions can also be seen in linear Markov games (Xie et al., 2020). But, again, as we shall see in Section 3.1, the imperfect information in POMGs brings significant difficulty in utilizing the linear structure over the reward functions compared with its fully observable counterpart. We also note that the regularity assumption imposed over $\boldsymbol{\phi}(\cdot,\cdot,\cdot)$ and $\boldsymbol{\theta}_h$ 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 & Szepesvári, 2020).

Regret Minimization For any product policy (μ, ν), the value function of (μ, ν) is defined as

$$V^{\mu,\nu} = \mathbb{E}\left[\sum_{h=1}^{H} r_h(s_h, a_h, b_h)\middle|\mu, \nu, \mathbb{P}\right]\,, \tag{2}$$

where the expectation is taken over the randomness of the underlying state transitions and the policies of both players. In this paper, we consider the learning objective of regret minimization. Without loss of generality, 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 $\{\mu_k\}_{k=1}^{t-1}$) before episode t. In specific, 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, defined as

$$\mathfrak{R}_{\max}^{T} = \max_{\mu^\dagger\in\Pi_{\max}} \mathbb{E}\left[\sum_{t=1}^{T}\left(V^{\mu^\dagger,\nu^t} - V^{\mu^t,\nu^t}\right)\right]\,. \tag{3}$$

In this work, we consider the regret minimization for the max-player in the *offline* setting, in which the max-player has access to the feature vectors of state-action weighted by min-player's policy ν^t in episode t (as well as transitions) before the t-th episode starts [3]. Note that this is slightly more general than the "offline" setting (also called *self-play*) considered by Chen et al. (2022); Xie et al. (2020), as we neither require the policy ν^t to be accessible to the max-player nor require both players to be directly controlled by a central controller.

Additional Notations With sequence-form representations, for any $\mu \in \Pi_{\max}$ and a sequence of functions $f = (f_h)_{h\in[H]}$ with $f_h : \mathcal{X}_h\times\mathcal{A} \to \mathbb{R}$, we let $\langle \mu, f\rangle \coloneqq \sum_{h\in[H]}\sum_{x_h\in\mathcal{X}_h, a\in\mathcal{A}} \mu_{1:h}(x_h, a_h) f_h(x_h, a_h)$. 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 hides all the logarithmic factors.

[3] Our second algorithm can work in a more general case where the max-player only receives such features after the t-th episode ends.

3 LEAST-SQUARES ONLINE MIRROR DESCENT

In this section, we present our LSOMD algorithm, as well as its theoretical guarantees.

3.1 LINEAR 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):

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

Hence, the regret in Eq. (3) can be rewritten as $\mathfrak{R}^T_{\max} = \max_{\mu^\dagger\in\Pi_{\max}} \sum_{t=1}^T \left\langle \mu^t - \mu^\dagger, \ell^t \right\rangle$, where we define the *loss function* in round t as

$$\ell^t_h(x_h,a_h) \coloneqq -\sum_{s_h\in x_h, b_h\in\mathcal{B}} p_{1:h}(s_h)\, \nu^t_{1:h}(y(s_h),b_h)\, \bar{r}_h(s_h,a_h,b_h)\ . \tag{4}$$

This implies that one can translate the regret minimization problem in Eq. (3) into a linear one.

To utilize the linear structure over the reward function to learn the unknown parameter $\boldsymbol{\theta}_h$, one may construct some sort of "linear" loss estimator $\hat{\boldsymbol{\theta}}_h$ of $\boldsymbol{\theta}_h$. However, this is more challenging in our case than it is in the case of linear bandits (Abbasi-Yadkori et al., 2011), linear MDPs (Jin et al., 2020), and linear perfect-information MGs (Xie et al., 2020), as we do 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 using the "feature vector" $\boldsymbol{\phi}(x_h,a_h)$ of (x_h,a_h), which is a composite feature vector weighted by opponent's policy ν and transition:

$$\boldsymbol{\phi}^{\nu^t}(x_h,a_h) \coloneqq -\sum_{(s_h,b_h)\in x_h\times\mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h),b_h)\boldsymbol{\phi}(s_h,a_h,b_h)\,, \tag{5}$$

which is assumed to be revealed to the max-player after the t-th episode ends in the offline setting as described in Section 2. Indeed, one can see that $\ell^t_h(x_h,a_h)$ is linear with $\boldsymbol{\phi}^{\nu^t}(x_h,a_h)$ and $\boldsymbol{\theta}_h$:

$$\ell^t_h(x_h,a_h) = \left\langle -\sum_{(s_h,b_h)\in x_h\times\mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h),b_h)\boldsymbol{\phi}(s_h,a_h,b_h), \boldsymbol{\theta}_h \right\rangle = \left\langle \boldsymbol{\phi}^{\nu^t}(x_h,a_h), \boldsymbol{\theta}_h \right\rangle .$$

Based on $\boldsymbol{\phi}^{\nu^t}(x_h,a_h)$, we further define the least-squares loss estimator $\hat{\boldsymbol{\theta}}_h$ as

$$\hat{\boldsymbol{\theta}}^t_h = \boldsymbol{Q}^{-1}_{\mu^t,h}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)r_h(s_h,a_h,b_h)\,, \tag{6}$$

where $\boldsymbol{Q}_{\mu^t,h} = \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu^t_{1:h}(x_h,a_h)\,\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top$ is the feature covariance matrix. Intuitively, this feature covariance matrix shares a similar spirit as its counterpart in the adversarial linear bandit literature (Lattimore & Szepesvári, 2020). However, we note that $\mu^t_{1:h}(\cdot,\cdot)$ here is not necessarily a distribution over $\mathcal{X}_h\times\mathcal{A}$.

This lemma shows $\hat{\boldsymbol{\theta}}^t_h$ is unbiased, which is critical in our analysis. See Appendix B.1 for its proof.

Lemma 3.1. *For any $t\in[T]$ and $h\in[H]$, it holds that $\mathbb{E}^{t-1}\left[\hat{\boldsymbol{\theta}}^t_h\right] = \boldsymbol{\theta}_h$.*

3.2 ALGORITHM DESCRIPTION

Our LSOMD algorithm follows the common scheme of OMD framework in that it runs OMD over $\Pi_{\max}$. Particularly, after interacting with the min-player using μ^t, it computes the loss estimate $\hat{\ell}^t_h(x_h,a_h)$ with $\hat{\boldsymbol{\theta}}^t_h$ defined in Eq. (6) (Line 6 - Line 10). Subsequently, it updates the policy $\hat{\mu}^{t+1}$ by solving a regularized linear optimization problem:

$$\hat{\mu}^{t+1} = \operatorname*{arg\,min}_{\mu\in\Pi_{\max}} \eta\left\langle \mu, \hat{\ell}^t \right\rangle + D_\Psi(\mu\|\hat{\mu}^t)\,, \tag{7}$$

Algorithm 1 LSOMD (max-player version)

1: **Input:** Tree-like structure of $\mathcal{X} \times \mathcal{A}$; Learning rate η.
2: **for** $t = 1$ to T **do**
3: **for** $h = 1$ to H **do**
4: Observe infoset x_h^t, execute $a_h^t \sim \mu_h^t(\cdot|x_h^t)$ and receive reward $r_h^t(s_h^t, a_h^t, b_h^t)$.
5: **end for**
6: **for** $h = 1$ to H **do**
7: Compute $\boldsymbol{Q}_{\mu^t,h} = \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}^t(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top$,
8: Compute $\hat{\boldsymbol{\theta}}_h^t = \boldsymbol{Q}_{\mu^t,h}^{-1}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)r_h(s_h,a_h,b_h)$,
9: **end for**
10: Construct loss estimate for all (x_h, a_h) and $h \in [H]$: $\hat{\ell}_h^t(x_h,a_h) = \langle \boldsymbol{\phi}^{\nu^t}(x_h,a_h), \hat{\boldsymbol{\theta}}_h \rangle$.
11: Receive composite features $\{\boldsymbol{\phi}^{\nu^{t+1}}(x,a)\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$.
12: Update policy: $\mu^{t+1} = (1-\gamma)\hat{\mu}^{t+1} + \gamma\pi$ with $\hat{\mu}^{t+1}$ computed in Eq. (7).
13: **end for**

where the potential function Ψ is chosen as $\Psi(\mu) = \sum_{h=1}^H \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h,a_h) \log\left(\frac{\mu_{1:h}(x_h,a_h)}{\sum_{a_h'\in\mathcal{A}}\mu_{1:h}(x_h,a_h')}\right)$. The induced Bregman divergence by Ψ is typically called *dilated* entropy distance-generating function and is also adopted by Hoda et al. (2010b); Kroer et al. (2015b); Kozuno et al. (2021). Moreover, we note that the optimization problem in Eq. (7) can be efficiently solved via a backward update detailed in Appendix C.1, which turns out to be an extension of it in the tabular case considered by Kozuno et al. (2021) to our linear case.

However, there is one more caveat. In the analysis of LSOMD, it is required to control the variance of the loss estimates. To this end, at the end of episode t, after receiving composite feature vectors $\{\boldsymbol{\phi}^{\nu^{t+1}}(x,a)\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$, our LSOMD algorithm will compute $\hat{\mu}^{t+1}$ by solving Eq. (7) and then mix it with a uniform policy π, *i.e.*, $\mu^{t+1} = (1-\gamma)\hat{\mu}^{t+1} + \gamma\pi$ and $\pi(a \mid x) = 1/A$ for any $(x,a) \in \mathcal{X}\times\mathcal{A}$, where μ^{t+1} is the policy to be used in the next episode and $\gamma \in (0,1)$ is the exploration parameter (Line 12).

3.3 Analysis

Due to leveraging the feature vectors of infoset-actions, we additionally require the following assumption, which essentially guarantees that each direction of the feature space is well explored by the uniform policy π.

Assumption 3.2. *The uniform policy π satisfies $\lambda_{\min}(\boldsymbol{Q}_{\pi,h}) \geq \rho > 0$, for any $h \in [H]$.*

The following theorem guarantees the regret upper bound of our LSOMD algorithm. Please see Appendix D for its proof.

Theorem 3.3. *In POMGs with linearly realizable rewards, by setting learning rate $\eta = \sqrt{\frac{\log A}{2TH(d+\rho^{-1})}}$ and exploration parameter $\gamma = \sqrt{\frac{X^2 \log A}{2HT(1+d\rho)\rho}}$, the regret bound of LSOMD in the offline setting is upper bounded by $\mathfrak{R}_{\max}^T \leq \mathcal{O}(\sqrt{(d+1/\rho)HTX^2\log A})$.*

Remark 3.4. *Compared with the regret upper bounds by Kozuno et al. (2021); Bai et al. (2022); Fiegel et al. (2023), the regret upper bound of our LSOMD does not have dependence on A, improves over Kozuno et al. (2021) by $\widetilde{\mathcal{O}}(\sqrt{HA})$ (omitting the dependence on d) but has an additional $\widetilde{\mathcal{O}}(\sqrt{HX})$ dependence compared with the minimax optimal result by Fiegel et al. (2023). On the other hand, as opposed to the high-probability regret guarantees in previous works studying tabular POMGs (Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023), the regret guarantee of our LSOMD algorithm only holds in expectation, which currently is not sufficient to be turned into an PAC algorithm for learning ε-NE. However, we would like to note again that this is the first line of algorithms that learns POMGs in the linear function approximation setting, with a regret guarantee independent of A. Also, we believe that it is possible to extend our results to high-probability results using self-concordant barrier potential functions and an increasing learning rate (Lee et al., 2020), which we leave as our future study.*

Technique Overview The proof of the regret upper bound of our LSOMD algorithm follows the common regret decomposition by bounding the *penalty* term and the *stability* term respectively. However, we note that bounding the stability term in our case is more difficult since bounding this term in the tabular case critically relies on the sparsity of the importance-weighted loss estimates, *i.e.*, the loss estimates are only non-zero at the experienced infoset-actions $\{(x_h^t, a_h^t)\}_{h\in[H]}$ (Kozuno et al., 2021). However, this does not apply in our case, where the linear loss estimator is utilized. To this end, we prove that the stability term in each episode t is (approximately) bounded by the summation of the log-partition function $\log Z_1^t(x_1)$ for all $x_1 \in \mathcal{X}_1$. We then bound this term by relating it with the expectation of the inner product between z^t and the loss estimate $\hat{\ell}^t$, in which $z_{1:h}^t(x_h, a_h)$ is the a_h-th value of the element-wise product of the random vectors independently sampled from categorical distributions specified by $\hat{\mu}^t$ along the path that leads to x_h (*cf.*, Appendix D.2). Also, the solutions to the update for OMD in previous works (Kozuno et al., 2021; Bai et al., 2022) are tailored to the tabular case and do not go through our problem, which we addressed by devising an efficient update for the linear case (*cf.*, Appendix C.1).

4 Least-Squares Follow-the-Regularized-Leader

This section presents the other FTRL-based algorithm, termed as LSFTRL, and its regret guarantees.

4.1 Algorithm Description

Our second LSFTRL algorithm uses the same linear loss estimates as LSOMD (Line 7 - Line 12). To update the policy μ^{t+1} used at episode $t+1$, it computes a linear optimization problem regularized by potential function $\{\Psi_h\}_{h\in[H}$ (Line 14):

$$\mu^{t+1} = \underset{\mu\in\Pi_{\max}}{\arg\min} \left\langle \mu, \hat{L}^t \right\rangle + \frac{1}{\eta} \sum_{h=1}^{H} \Psi_h \left(p_{1:h}^\star \cdot \mu_{1:h}\right) , \tag{8}$$

where $\hat{L}^t = \sum_{k=1}^t \hat{\ell}^k$ is the cumulative loss estimate, $p_{1:h}^\star(x_h) = p_0^\star(x_1) \prod_{h'=1}^{h-1} p_{h'}^\star(x_{h'+1}|x_{h'}, a_{h'})$ with $p_h^\star(\cdot|x_h, a_h) \in \Delta_{C(x_h,a_h)}$ being a kind of "transition probability function" over $\mathcal{X}_h \times \mathcal{A} \times \mathcal{X}_{h+1}$, and $p_{1:h}^\star \cdot \mu_{1:h}$ is defined as $[p_{1:h}^\star \cdot \mu_{1:h}](x_h, a_h) = p_{1:h}^\star(x_h)\mu_{1:h}(x_h, a_h)$. Note that such $p^\star$ is well-defined due to the perfect recall condition, and $p_{1:h}^\star \cdot \mu_{1:h}$ is a probability distribution over the infoset-action pair $\mathcal{X}_h \times \mathcal{A}$ at step h. We also remark that similar approaches that combine the FTRL/OMD with $p_{1:h}^\star(\cdot)$ have also been exploited in previous works (*e.g.*, the balanced transition $p^\star$ of Bai et al. (2022); Fiegel et al. (2023) and the adversarial transition $p^{\star,\nu^t}$ of Fiegel et al. (2023)), but we will choose a different $p_{1:h}^\star(\cdot)$ satisfying

$$p^\star = \underset{\tilde{p}\in\mathbb{P}^\star}{\arg\max} \min_{h\in[H], x_h\in\mathcal{X}_h} \tilde{p}_{1:h}(x_h) , \tag{9}$$

where $\mathbb{P}^\star$ denotes the set of all the valid transitions over infoset-actions. The computation of such $p^\star$ can be efficiently implemented using backward dynamic programming in $\mathcal{O}(XA)$ time, the details of which are postponed to Appendix E.2.3. As we shall see, the property of such $p^\star$ will serve as a key ingredient of the regret upper bound of our LSFTRL algorithm. Besides, LSFTRL chooses $\Psi_h(w_h) = \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} w_h(x_h, a_h) \log(w_h(x_h, a_h))$ as the negative entropy potential function (not to be confused with the dilated entropy potential function used in LSOMD). We also note that the computation of Eq. (8) can also be efficiently solved by reducing the update of LSFTRL to an OMD-like update, the details of which are deferred to Appendix C.3. The complete pseudo-code for LSFTRL algorithm is postponed to Appendix C.2.

4.2 Analysis

Let $p_{1:h}^\nu(x_h) = \sum_{s_h\in x_h} p_{1:h}(s_h)\, \nu_{1:h-1}(y(s_{h-1}), b_{h-1})$, 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. Similar to LSOMD, the regret upper bound of LSFTRL also depends on an exploratory assumption, detailed in the following. Please see Appendix E.1 for additional discussions on this assumption.

Assumption 4.1. *For any $t \in [T]$, $h \in [H]$ and $x_1, x_2 \in \mathcal{X}_h$, it holds that $p_{1:h}^{\nu^t}(x_1)/p_{1:h}^\star(x_2) \le \lambda$.*

We now present the regret upper bound of LSFTRL, with its proof postponed to Appendix E.2.

Theorem 4.2. *In POMGs with linearly realizable rewards, by setting learning rate* $\eta = \sqrt{\frac{2\log(AX)}{Td\lambda}}$, *the regret bound of* LSFTRL *in the offline setting is upper bounded by* $\mathfrak{R}^T_{\max} \leq \widetilde{\mathcal{O}}(\sqrt{H^2 d\lambda T})$.

Remark 4.3. LSFTRL *obtains the regret guarantee, which eliminates the dependence on both* X *and* A, *in exchange for an exploratory assumption depending on the opponent's policy* ν^t. *Compared with previous results, the regret upper bound in Theorem 4.2 improves over the minimax optimal regret* $\widetilde{\mathcal{O}}(\sqrt{XAT})$ *of Fiegel et al. (2023) by a factor* $\widetilde{\mathcal{O}}(\sqrt{XA/H^2})$ *(omitting the dependence on* d *and* λ*). Note that if the max-player does not have access to* λ, *we can instead set* $\eta = \sqrt{2\log(AX)/(Td)}$ *without requiring the knowledge of* λ, *but at a slight cost of having the regret changing from* $\widetilde{\mathcal{O}}(\sqrt{H^2 d\lambda T})$ *to* $\widetilde{\mathcal{O}}(\lambda\sqrt{H^2 dT})$. *Besides, in cases where* λ *is undesirably large (e.g.,* $\lambda \geq X/H$*), a different choice of* $p^\star$ *by setting* $p^\star_{1:h}(x_h) \equiv 1$ *leads to the following regret guarantee of* LSFTRL.

Theorem 4.4. *In POMGs with linearly realizable rewards, by setting learning rate* $\eta = \sqrt{\frac{2X\log A}{THd}}$ *and* $p^\star_{1:h}(x_h) \equiv 1$ *for any* $x_h \in \mathcal{X}$ *and* $h \in [H]$, *the regret bound of* LSFTRL *in the offline setting is upper bounded by* $\mathfrak{R}^T_{\max} \leq \widetilde{\mathcal{O}}(\sqrt{HXdT})$.

Remark 4.5. *The proof of Theorem 4.4 is deferred to Appendix E.3. Note that by setting* $p^\star_{1:h}(x_h) \equiv 1$, $p^\star$ *is no longer a transition function over infoset-actions. Importantly, the regret in Theorem 4.4 improves over the minimax optimal result of Fiegel et al. (2023) by a factor* $\widetilde{\mathcal{O}}(\sqrt{A/H})$ *(omitting the dependence on* d*).*

Technique Overview We bound the regret of LSFTRL also by decomposing the regret into the penalty term and the stability term (Lattimore & Szepesvári, 2020), which is also adopted by Fiegel et al. (2023). However, we bound the stability term of LSFTRL with particular care such that the variances of the loss estimates are well-controlled by λ in Assumption 4.1 and d (*cf.*, Appendix E.2). Moreover, when bounding the penalty of LSFTRL with $p^\star_{1:h}(x_h) = 1$, we establish a refined analysis that shaves off an $\mathcal{O}(\sqrt{A})$ factor, compared with the direct combination of the original analysis of Fiegel et al. (2023) and the setting of $p^\star_{1:h}(x_h) = 1$ (*cf.*, Appendix E.3).

4.3 Regret Lower Bound

We also provide a regret lower bound of learning POMGs with linearly realizable rewards in the following theorem, the proof of which is deferred to Appendix F.

Theorem 4.6. *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 POMG instance on which* $\mathfrak{R}^T_{\max} \geq \Omega(\sqrt{d\min(d,H)T})$.

Remark 4.7. *We conjecture that the regret lower bound can be further improved to* $\mathfrak{R}^T_{\max} \geq \Omega(\sqrt{dHT})$, *and currently our regret upper bounds of* LSOMD *and* LSFTRL *with the second initialization are loose by* $\widetilde{\mathcal{O}}(X)$ *and* $\widetilde{\mathcal{O}}(\sqrt{X})$ *factors and regret upper bound of* LSFTRL *with the first initialization is loose by an* $\widetilde{\mathcal{O}}(\sqrt{H})$ *factor (omitting the dependence on* ρ *and* λ*). We leave the investigation into the possible improvements of the upper and lower bounds as our future studies.*

5 Conclusion

In this work, we make the first step towards provably efficient learning of the two-player, zero-sum IIEFGs with linear function approximation, in the formulation of POMGs with linearly realizable rewards and unknown transitions. It is proven that, the proposed LSOMD algorithm obtains an $\widetilde{\mathcal{O}}(\sqrt{(d+1/\rho)HX^2T})$ regret, and the LSFTRL algorithm attains regret of orders $\widetilde{\mathcal{O}}(\sqrt{H^2 d\lambda T})$ and $\widetilde{\mathcal{O}}(\sqrt{HXdT})$. We accomplish this by devising the first least-squares loss estimator for this setting, along with new ingredients in the analysis for both the LSOMD and LSFTRL algorithms, which may be of independent interest. Also, we provide an $\Omega(\sqrt{d\min(d,H)T})$ regret lower bound. Besides, there are also several interesting future directions to be explored. One natural question might be how to obtain high-probability results in this challenging problem so as to find an ε-NE. The other question might be whether it is possible generalize the proposed algorithms and results to multi-player general-sum POMGs. We hope our results may shed light on better understandings of learning large-scale POMGs and we leave these extensions as our further studies.

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. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 2312–2320, 2011.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 463–474. PMLR, 2020.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 551–560. PMLR, 2020.

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, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1337–1382. PMLR, 2022.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 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, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 793–802. PMLR, 2019.

Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting CFR+ and alternating updates. *J. Artif. Intell. Res.*, 64:429–443, 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, 29 March - 1 April 2022, Paris, France*, volume 167 of *Proceedings of Machine Learning Research*, pp. 227–261. PMLR, 2022.

Qiwen Cui and Lin F. Yang. Minimax sample complexity for turn-based stochastic game. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1496–1504. AUAI Press, 2021.

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, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2651–2652. PMLR, 2023.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control, L4DC 2020, Online Event, Berkeley, CA, USA, 11-12 June 2020*, volume 120 of *Proceedings of Machine Learning Research*, pp. 486–489. PMLR, 2020.

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, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 5381–5390. AAAI Press, 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, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3018–3028. PMLR, 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, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 5372–5380. AAAI Press, 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, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10093–10135. 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, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Amy Greenwald and Keith Hall. Correlated q-learning. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 242–249. AAAI Press, 2003.

Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J. ACM*, 60(1): 1:1–1:16, 2013.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12790–12822. PMLR, 2023.

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, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 805–813. JMLR.org, 2015.

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Math. Oper. Res.*, 35(2):494–512, 2010a.

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing Techniques for Computing Nash Equilibria of Sequential Games. *Mathematics of Operations Research*, 2010b.

Junling Hu and Michael P. Wellman. Nash q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4:1039–1069, 2003.

Zeyu Jia, Lin F. Yang, and Mengdi Wang. Feature-based q-learning for two-player stochastic games. *CoRR*, abs/1906.00423, 2019.

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, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. 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, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10251–10279. 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, Valencia, Spain, June 4-8, 2012 (3 Volumes)*, pp. 837–846. IFAAMAS, 2012.

Daphne Koller and Nimrod Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and economic behavior*, 4(4):528–552, 1992.

Daphne Koller, Nimrod Megiddo, and Bernhard Von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and economic behavior*, 14(2):247–259, 1996.

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, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 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 '15, Portland, OR, USA, June 15-19, 2015*, pp. 817–834. ACM, 2015a.

Christian Kroer, Kevin Waugh, Fatma Kilinç-Karzan, and Tuomas Sandholm. Faster First-Order Methods for Extensive-Form Game Solving . In *ACM Conference on Economics and Computation*, pp. 817–834, 2015b.

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, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 872–882, 2018.

HW Kuhn. Extensive games and the problem of information. *Contributions to the Theory of Games*, (24):193, 1953.

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*, 13 (1):99–110, 2021.

Michail G. Lagoudakis and Ronald Parr. Value function approximation in zero-sum markov games. In *UAI '02, Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence, University of Alberta, Edmonton, Alberta, Canada, August 1-4, 2002*, pp. 283–292. Morgan Kaufmann, 2002.

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. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pp. 1078–1086. Curran Associates, Inc., 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, NeurIPS 2020, December 6-12, 2020, virtual*, 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, December 6-14, 2021, virtual*, pp. 14293–14305, 2021.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent RL in markov games with a generative model. In *NeurIPS*, 2022.

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. *CoRR*, abs/2003.13590, 2020.

Michael L. Littman. Friend-or-foe q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pp. 322–328. Morgan Kaufmann, 2001.

Michael L. Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996*, pp. 310–318. Morgan Kaufmann, 1996.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7001–7010. PMLR, 2021.

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, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13717–13745. PMLR, 2022.

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*, 356(6337):508–513, 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, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7119–7129. PMLR, 2020.

John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

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, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Julien Pérolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1321–1329. JMLR.org, 2015.

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, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 2157–2164. AAAI Press, 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. *CoRR*, abs/2112.03178, 2021.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5192–5202, 2018.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Oskari Tammelin. Solving large imperfect information games using CFR+. *CoRR*, 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, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 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, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2793–2848. 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, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3674–3682. PMLR, 2020.

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, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24496–24523. 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, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40804–40829. PMLR, 2023.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12653–12662. PMLR, 2021.

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, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4532–4576. 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*, pp. 467–475, 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, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 1729–1736. Curran Associates, Inc., 2007.

A PROPERTIES OF THE GAME

The lemma below delineates the key property of $p^\star$ as transition probability functions.

Lemma A.1. *For any $h \in [H]$, any $p^\star$ as transition probability function 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 functions over infoset-actions, we have

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

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

B PROPERTIES OF THE LEAST-SQUARES LOSS ESTIMATOR

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

B.1 UNBIASNESS OF THE LOSS ESTIMATOR

Proof of Lemma 3.1.

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

□

B.2 VARIANCE OF THE LOSS ESTIMATOR

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

Lemma B.1. *For any $h \in [H]$, it holds that*

$$\mathbb{E}^{t-1}\left[\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu^t_{1:h}(x_h,a_h)\hat{\ell}^t_h(x_h,a_h)^2\right] \le d\,. \tag{10}$$

Proof.

$$
\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\left(\hat{\boldsymbol{\theta}}^t_h\right)^\top\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\\
&\qquad\cdot\mathbb{E}^{\mu^t,\nu^t}\left[r_h(s_h,a_h,b_h)^2\boldsymbol{Q}^{-1}_{\mu^t,h}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\boldsymbol{Q}^{-1}_{\mu^t,h}\right]\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\\
&\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}^{-1}_{\mu^t,h}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top\boldsymbol{Q}^{-1}_{\mu^t,h}\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}^{-1}_{\mu^t,h}\\
&\qquad\cdot\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}p^{\nu^t}(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}^{-1}_{\mu^t,h}\boldsymbol{\phi}^{\nu^t}(x_h,a_h)\\
&=\mathrm{tr}\left[\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}^{-1}_{\mu^t,h}\right)\right.\\
&\qquad\left.\cdot\left(\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}p^{\nu^t}(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}^{-1}_{\mu^t,h}\right)\right]\\
&=\mathrm{tr}\left[\boldsymbol{I}_d\cdot\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}p^{\nu^t}(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}^{-1}_{\mu^t,h}\right]\\
&\le\mathrm{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}^{-1}_{\mu^t,h}\right]\\
&=\mathrm{tr}\left[\boldsymbol{I}_d\right]=d\,.
\end{aligned}
$$

□

C Computation Issue

In this section, we present efficient solutions to the optimization problems of both LSOMD and LSFTRL.

C.1 Efficient Update for LSOMD

To begin with, we first introduce a generalized version of OMD update in Eq. (7), which leverages learning rates adaptive to each infoset. Specifically, given any list of learning rates $\eta := (\eta_h(x_h))_{h\in[H],x_h\in\mathcal{X}}$, the potential function is defined as

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

By the fact that for all positive $\mu\in\Pi_{\max}$, 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 dilated distance generating function

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

The generalized version of OMD update in Eq. (7) is given as follows:

$$\begin{aligned}\mu^{t+1} &= \arg\min_{\mu\in\Pi_{\max}} \left\langle \mu, \hat{\ell}^t \right\rangle + D_{\Psi_\eta}(\mu\|\mu^t) \\ &= \arg\min_{\mu\in\Pi_{\max}} \left\langle \mu, \hat{\ell}^t \right\rangle + \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}(x_h,a_h)}{\eta_h(x_h)} \log \frac{\mu_h(a_h|x_h)}{\mu^t_h(a_h|x_h)} . \end{aligned} \tag{11}$$

We remark that $\eta := (\eta_h(x_h))_{h\in[H],x_h\in\mathcal{X}}$ also generalizes the of *balanced transitions* used in Farina et al. (2020); Bai et al. (2022); Fiegel et al. (2023).

The solution to Eq. (11) is given in the following proposition. Notice that the solution to this optimization problem 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. However, our solution supports the loss estimator with non-zero loss estimates for arbitrary infoset-action pairs.

Proposition C.1. *The solution to the update rule in Eq. (11) is as followed:*

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

where

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

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

Proof. We first note that

$$\begin{aligned} &\left\langle \mu, \hat{\ell}^t \right\rangle + D_{\Psi_\eta}(\mu\|\mu^t) \\ &= \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h,a_h) \left[\hat{\ell}^t_h(x_h,a_h) + \frac{1}{\eta_h(x_h)} \log \frac{\mu_h(a_h|x_h)}{\mu^t_h(a_h|x_h)} \right] \\ &= \sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_h} \mu_{1:h-1}(x_h) \left[\left\langle \mu_h(\cdot|x_h), \hat{\ell}^t_h(x_h,\cdot) \right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_h(\cdot|x_h)||\mu^t_h(\cdot|x_h)\right)}{\eta_h(x_h)} \right] . \end{aligned} \tag{12}$$

We now prove the proposition through backward induction over $h = H,\dots,1$. For $h = H, x_H \in \mathcal{X}_H$, it is easy to see that

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

where $Z^t_H(x_H) = \sum_{a_H\in\mathcal{A}} \mu^t_H(a_H|x_H) \exp\left\{ -\eta_h(x_H)\hat{\ell}^t_H(x_H,a_H) \right\} > 0$ is a normalization factor.

Suppose the proposition holds from step $h+1$ to H and consider the h-th step. Substituting the induction hypothesis, one can see that Eq. (12) can be expressed as follows:

$$\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}^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] \\
&= \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] \\
&\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^t_{h'+1}(x_{h'+1}) - \sum_{x_{h'}\in\mathcal{X}_{h'}} \frac{\mu_{1:h'-1}(x_{h'})}{\eta_{h'}(x_{h'})} \log Z^t_{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] \\
&\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^t_{h+1}(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}^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] \\
&\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}^t_h(x_h,\cdot) - \sum_{x_{h+1}\in C(x_h,\cdot)} \frac{\log Z^t_{h+1}(x_{h+1})}{\eta_{h+1}(x_{h+1})}\right\rangle + \frac{D_{\mathrm{KL}}\left(\mu_h(\cdot|x_h)||\mu^t_h(\cdot|x_h)\right)}{\eta_h(x_h)}}_{\heartsuit}\Bigg].
\end{aligned}$$

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

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

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

which thus concludes the proof. □

Proposition C.1 immediately implies the efficient update procedure for LSOMD, detailed in Algorithm 2, by setting $\eta_h(x_h) \equiv \eta$ for all $x_h \in \mathcal{X}$ in Proposition C.1. In what follows, for notational convenience, we denote $J^t_h(x_h,a_h) = -\eta_h(x_h)\hat{\ell}^t_h(x_h,a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h(x_h)}{\eta_{h+1}(x_{h+1})} \log Z^t_{h+1}(x_{h+1})$ as the surrogate loss.

C.2 LSFTRL Algorithm

C.3 Efficient Update for LSFTRL

To solve the update of LSFTRL, we follow the same idea as Fiegel et al. (2023) that translating the update of FTRL into the update of OMD-like update. In specific, the Proposition F.2 of Fiegel et al. (2023) shows that the update of Eq. (8) is equivalent to the solution to the following optimization problem:

$$\mu^t = \arg\min_{\mu\in\Pi_{\max}} \left\langle \mu, \hat{L}^t\right\rangle + D_{\eta^\star}(\mu,\mu^\star), \tag{13}$$

Algorithm 2 Update-of-LSOMD

Input: Tree-like structure of $\mathcal{X} \times \mathcal{A}$, $\hat{\mu}^t$ given by update Eq. (7); fixed learning rates η; the loss estimates $\left\{\hat{\ell}_h^t(x_h, a_h)\right\}_{(x_h,a_h)\in\mathcal{X}\times\mathcal{A}}$.
Initialization: For all x_H in $\mathcal{X}_H$, initialize $Z^t(x_{H+1}) = 1$.
for $h = H$ to 1 **do**
  for x_h in $\mathcal{X}_h$ **do**
    Compute $J_h^t(x_h, a_h) = -\eta\hat{\ell}_h^t(x_h, a_h) + \sum_{x_{h+1}\in C(x_h,a_h)} \log Z_{h+1}^t(x_{h+1})$,
    Compute $Z_h^t(x_h) = \sum_{a_h\in\mathcal{A}} \hat{\mu}_h^t(a_h|x_h) \exp\left(J_h^t(x_h, a_h)\right)$,
    for a_h in $\mathcal{A}$ **do**
      Compute $\hat{\mu}_h^{t+1}(a_h|x_h) = \hat{\mu}_h^t(a_h|x_h) \exp\left(J_h^t(x_h, a_h) - \log Z_h^t(x_h)\right)$.
    end for
  end for
end for

Algorithm 3 LSFTRL (max-player version)

Input: Tree-like structure of $\mathcal{X} \times \mathcal{A}$; Learning rates η; $p^\star$.
for $t = 1$ to T **do**
  for $h = 1$ to H **do**
    Observe infoset x_h^t.
    Execute $a_h^t \sim \mu_h^t(\cdot|x_h^t)$ and receive reward $r_h^t(s_h^t, a_h^t, b_h^t)$.
  end for
  Receive composite features $\left\{\boldsymbol{\phi}^{\nu^t}(x, a)\right\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$.
  for $h = 1$ to H **do**
    Compute $\boldsymbol{Q}_{\mu^t,h} = \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}^t(x_h, a_h)\boldsymbol{\phi}^{\nu^t}(x_h, a_h)\boldsymbol{\phi}^{\nu^t}(x_h, a_h)^\top$,
    Compute $\hat{\boldsymbol{\theta}}_h^t = \boldsymbol{Q}_{\mu^t,h}^{-1}\boldsymbol{\phi}^{\nu^t}(x_h, a_h)r_h(s_h, a_h, b_h)$,
  end for
  Construct loss estimate for all (x_h, a_h) and $h \in [H]$: $\hat{\ell}_h^t(x_h, a_h) = \left\langle \boldsymbol{\phi}^{\nu^t}(x_h, a_h), \hat{\boldsymbol{\theta}}_h \right\rangle$.
  Compute cumulative loss estimate at episode t: $\hat{L}^t = \hat{L}^{t-1} + \hat{\ell}^t$.
  Compute update (8) using Update-of-LSFTRL.
end for

where $\eta^\star := (\eta_h^\star(x_h))_{h,x_h}$ is a learning rate adaptive to each infoset, $\mu^\star$ is a base policy and we define $D_{\eta^\star}\left(\mu^1, \mu^0\right) := \sum_{h=1}^H \sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \frac{\mu_{1:h}^1(x_h,a_h)}{\eta_h^\star(x_h)} \log \frac{\mu_h^1(a_h|x_h)}{\mu_h^0(a_h|x_h)}$.

Therefore, to solve Eq. (8), for all $x_h \in \mathcal{X}$, we first set the adaptive learning rate $\eta^\star$ as

$$\eta_h^\star(x_h) = \frac{\eta}{(H - h + 1)p_{1:h}^\star(x_h)}, \tag{14}$$

and set the base policy $\mu^\star$ as

$$\mu^\star = \arg\min_{\mu^\star\in\Pi_{\max}} \sum_{h=1}^H \Psi_h\left(p_{1:h}^\star \cdot \mu_{1:h}^\star\right), \tag{15}$$

which can be computed efficiently via backward dynamic programming in $\mathcal{O}(XA)$ time. Then, combined with the efficient update procedure of LSOMD in Algorithm 2, the solution to the update of LSFTRL can be obtained by substituting μ^t with $\mu^\star$, the details of which are presented in Algorithm 4 for completeness.

D Proof of Regret Guarantee of LSOMD

In this section, we present the proof of the regret guarantee of LSOMD.

Algorithm 4 `Update-of-LSFTRL`

Input: Tree-like structure of $\mathcal{X} \times \mathcal{A}$; fixed learning rates η ; transition probability function $p^\star$; 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. (14) ; Set base policy $\mu^\star$ according to Eq. (15) .
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

Proof of Theorem 3.3. First note that

$$\begin{aligned}\left\langle \hat{\mu}^t - \mu^\dagger, \hat{\ell}^t \right\rangle &= D_\eta(\hat{\mu}^t\|\hat{\mu}^{t+1}) - D_\eta(\hat{\mu}^t\|\hat{\mu}^t) - (D_\eta(\mu^\dagger\|\hat{\mu}^{t+1}) - D_\eta(\mu^\dagger\|\hat{\mu}^t)) \\ &= D_\eta(\mu^\dagger\|\hat{\mu}^t) - D_\eta(\mu^\dagger\|\hat{\mu}^{t+1}) + D_\eta(\hat{\mu}^t\|\hat{\mu}^{t+1}) .\end{aligned}$$

Taking summation of the above display over t and telescoping the sum, we have

$$\sum_{t=1}^T \left\langle \hat{\mu}^t - \mu^\dagger, \hat{\ell}^t \right\rangle \le \underbrace{D_\eta(\mu^\dagger\|\hat{\mu}^1)}_{\textsc{Penalty}} + \underbrace{\sum_{t=1}^T D_\eta(\hat{\mu}^t\|\hat{\mu}^{t+1})}_{\textsc{Stability}} . \tag{16}$$

On the other hand, by the unbiasedness of $\hat{\ell}^t$ and the tower rule, it holds that

$$\mathbb{E}\left[\left\langle \mu^t - \mu^\dagger, \ell^t \right\rangle\right] = \mathbb{E}\left[\mathbb{E}^{t-1}\left[\left\langle \mu^t - \mu^\dagger, \hat{\ell}^t \right\rangle\right]\right] = \mathbb{E}\left[\left\langle \mu^t - \mu^\dagger, \hat{\ell}^t \right\rangle\right] , \tag{17}$$

where recall that $\mu^t = (1-\gamma)\hat{\mu}^t + \gamma\pi^t$.

Combining Eq. (16) and Eq. (17), along with the definition of regret in Eq. (3), one can deduce that

$$\begin{aligned}\mathfrak{R}_{\max}^T &\le \max_{\mu^\dagger\in\Pi_{\max}} (1-\gamma)\mathbb{E}\left[\underbrace{D_\eta(\mu^\dagger\|\hat{\mu}^1)}_{\textsc{Penalty}} + \underbrace{\sum_{t=1}^T D_\eta(\hat{\mu}^t\|\hat{\mu}^{t+1})}_{\textsc{Stability}}\right] + 2\gamma HT \\ &\le (1-\gamma)\left(\frac{X\log A}{\eta} + \eta TXHd\right) + 2\gamma HT \\ &\le \left(\frac{X\log A}{\eta} + \eta TXHd\right) + 2\eta HXTd\alpha^{-1} \\ &= \frac{X\log A}{\eta} + 2\eta TXHd(1+\alpha^{-1}) ,\end{aligned}$$

where the second inequality comes from Lemma D.1 and D.5.

Finally, the proof is concluded by substituting $\eta = \sqrt{\frac{\log A}{2THd(1+\alpha^{-1})}}$ and $\gamma = \sqrt{\frac{Xd\log A\alpha^{-1}}{2HT(1+\alpha^{-1})}}$. □

In our `LSOMD`, we set the learning rate to be constant, *i.e.*, $\eta_h(x_h) \equiv \eta$ for all $x_h \in \mathcal{X}$.

D.1 Bounding the Penalty Term

Lemma D.1. *The* Penalty *term is bounded by*

$$\textsc{Penalty} \leq \frac{X \log A}{\eta}.$$

Proof.

$$\begin{aligned}
D_\eta(\mu^\dagger \| \hat{\mu}^1) &= \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu^\dagger_{1:h}(x_h,a_h)}{\eta} \log \frac{\mu^\dagger_h(a_h|x_h)}{\hat{\mu}^1_h(a_h|x_h)} \\
&\leq \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu^\dagger_{1:h}(x_h,a_h)}{\eta} \log \hat{\mu}^1_h(a_h|x_h) \\
&= \log A \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu^\dagger_{1:h}(x_h,a_h)}{\eta} \\
&\leq \frac{X \log A}{\eta}.
\end{aligned}$$

□

D.2 Bounding the Stability Term

To begin with, we first introduce the following lemma, which is a generalized version of Lemma D.7 by Bai et al. (2022). Intuitively, this lemma states that the one-step stability term can be bounded by the inner product between $\hat{\mu}$ and $\hat{\ell}^t$ as well as the summation of log-partition function $\log Z_1^t$.

Lemma D.2. *For given η and any $\mu \in \Pi_{\max}$, we have*

$$D_\eta(\mu\|\hat{\mu}^{t+1}) - D_\eta(\mu\|\hat{\mu}^t) = \left\langle \mu, \hat{\ell}^t \right\rangle + \sum_{x_1\in\mathcal{X}_1} \frac{1}{\eta_1(x_1)} \log Z_1^t(x_1). \tag{18}$$

Proof.

$$\begin{aligned}
&D_\eta(\mu\|\hat{\mu}^{t+1}) - D_\eta(\mu\|\hat{\mu}^t) \\
&= \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}(x_h,a_h)}{\eta_h(x_h)} \log \frac{\hat{\mu}^t_h(a_h|x_h)}{\hat{\mu}^{t+1}_h(a_h|x_h)} \\
&= \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu^t_{1:h}(x_h,a_h)}{\eta_h(x_h)} \left(\eta_h(x_h)\hat{\ell}^t_h(x_h,a_h) - \sum_{x_{h+1}\in C(x_h,a_h)} \frac{\eta_h(x_h)}{\eta_{h+1}(x_{h+1})} \log Z^t_{h+1}(x_{h+1}) \right) \\
&\qquad + \sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_h} \frac{\mu_{1:h-1}(x_h)}{\eta_h(x_h)} \log Z^t_h(x_h) \\
&= \left\langle \mu, \hat{\ell}^t \right\rangle + \sum_{h=1}^{H} \left[- \sum_{x_{h+1}\in\mathcal{X}_{h+1}} \frac{\mu_{1:h}(x_{h+1})}{\eta_{h+1}(x_{h+1})} \log Z^t_{h+1}(x_{h+1}) + \sum_{x_h\in\mathcal{X}_h} \frac{\mu_{1:h-1}(x_h)}{\eta_h(x_h)} \log Z^t_h(x_h) \right] \\
&= \left\langle \mu, \hat{\ell}^t \right\rangle + \sum_{x_1\in\mathcal{X}_1} \frac{1}{\eta_1(x_1)} \log Z^t_1(x_1).
\end{aligned}$$

□

From Lemma D.2 and setting $\eta_h(x_h) \equiv \eta$, we have

$$D_\eta(\hat{\mu}^t \| \hat{\mu}^{t+1}) = \left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle + \sum_{x_1 \in \mathcal{X}_1} \frac{1}{\eta_1(x_1)} \log Z_1^t(x_1)$$
$$= \left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle + \frac{1}{\eta} \sum_{x_1 \in \mathcal{X}_1} \log Z_1^t(x_1) \,. \tag{19}$$

Hence, to bound the STABILITY term, it suffices to bound the log-partition function $\log Z_1^t$. To this end, roughly speaking, we prove that the summation of all $\log Z_1^t(x_1)$ for $x_1 \in \mathcal{X}_1$ can be bounded by the product between (a) the value of all the reachable (x_h, a_h) in the element-wise product of the random vectors independently sampled from the categorical distributions specified by $\hat{\mu}^t(\cdot|x_h)$; and (b) the loss estimate at (x_h, a_h). Compared with the analysis tailored to the importance-weighted loss estimate in previous works (Kozuno et al., 2021; Fiegel et al., 2023), where bounding similar log-partition function $\log Z_1^t(x_1)$ is easier and can be done by only considering the random variables sampled from the Bernoulli distributions along the experienced trajectory, our analysis for least-squares loss estimate is more challenging and also generalizes it in previous works.

D.2.1 BOUNDING THE LOG-PARTITION FUNCTION $\log Z_h^t$

We first define $\boldsymbol{z}^t(x_h, \cdot) \in \{0,1\}^A$, which is a random vector independently sampled from the categorical distribution parameterized by $\hat{\mu}_h^t(\cdot|x_h)$, by

$$\boldsymbol{z}^t(x_h, \cdot) \sim \mathrm{Cat}(\hat{\mu}_h^t(\cdot|x_h)) \,,$$

such that $\mathbb{P}\left(\boldsymbol{z}^t(x_h, a_h) = 1\right) = \hat{\mu}_h^t(a_h|x_h)$.

Notice that

$$\mathbb{E}\left[\prod_{h'=1}^{h} \boldsymbol{z}^t(x_{h'}, a_{h'})\right] = \hat{\mu}_{1:h}^t(x_h, a_h) \,.$$

We also let

$$\boldsymbol{z}^t{}_{h:h'}(x_{h'}, a_{h'}) = \prod_{h''=h}^{h'} \boldsymbol{z}^t(x_{h''}, a_{h''}) \,,$$

where $\{(x_{h''}, a_{h''})\}_{h'' \in [h,h']}$ is the unique path from (x_h, a_h) to $(x_{h'}, a_{h'})$ (under perfect recall condition). Besides, we denote the product of $Z_{h+1}^t(x_{h+1})$ for all $x_{h+1} \in C(x_h, a_h)$ as

$$\Xi^t(x_h, a_h) = \prod_{x_{h+1} \in C(x_h, a_h)} Z_{h+1}^t(x_{h+1}) \,,$$

so that

$$\frac{1}{\eta} \sum_{x_1 \in \mathcal{X}_1} \log Z_1^t(x_1) = \frac{1}{\eta} \log \Xi^t(\emptyset) \,.$$

Then, the following lemma shows that, $\Xi_h^t(x_h, a_h)$ is equivalent to the expectation of the exponentiation of the summation of $\boldsymbol{z}^t{}_{h+1:h'}(x_{h'}, a_{h'}) \hat{\ell}_{h'}^t(x_{h'}, a_{h'})$, where $(x_{h'}, a_{h'})$ are all the reachable infoset-action pairs from (x_h, a_h).

Lemma D.3. *For any $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$ and $h \in [H-1]$, we have*

$$\Xi_h^t(x_h, a_h) = \mathbb{E}_{\boldsymbol{z}^t}\left[\exp\left(-\eta \sum_{h'=h+1}^{H} \sum_{(x_{h'}, a_{h'}) \in C_{h'}(x_h, a_h)} \boldsymbol{z}^t{}_{h+1:h'}(x_{h'}, a_{h'}) \hat{\ell}_{h'}^t(x_{h'}, a_{h'})\right)\right] . \tag{20}$$

As an immediate corollary of Lemma D.3, we have

$$\Xi^t(\emptyset) = \mathbb{E}_{\boldsymbol{z}^t}\left[\exp\left(-\eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h)\right)\right] .$$

Proof. We prove this by backward induction. For $h = H - 1$, we have

$$\begin{aligned}\Xi^t_{H-1}(x_{H-1}, a_{H-1}) &= \prod_{x_H \in C(x_{H-1}, a_{H-1})} \sum_{a_H \in \mathcal{A}} \hat{\mu}^t_H(a_H|x_H) \exp(-\eta \hat{\ell}^t_H(x_H, a_H)) \\ &= \mathbb{E}_{\boldsymbol{z}^t}\left[\exp\left(-\eta \boldsymbol{z}^t{}_{H:H}(x_H, a_H)\hat{\ell}^t_H(x_H, a_H)\right)\right].\end{aligned}$$

Suppose Eq. (20) holds from $h' = h$ to H. Then for $h' = h - 1$, one can deduce that

$$\begin{aligned}&\Xi^t_{h-1}(x_{h-1}, a_{h-1}) \\ =& \prod_{x_h \in C(x_{h-1}, a_{h-1})} \sum_{a_h \in \mathcal{A}} \hat{\mu}^t_h(a_h|x_h) \exp\left(-\eta \hat{\ell}^t_h(x_h, a_h)\right) \Xi^t_h(x_h, a_h) \\ =& \prod_{x_h \in C(x_{h-1}, a_{h-1})} \sum_{a_h \in \mathcal{A}} \hat{\mu}^t_h(a_h|x_h) \exp\left(-\eta \hat{\ell}^t_h(x_h, a_h)\right) \\ &\quad \cdot \mathbb{E}_{\boldsymbol{z}^t}\left[\exp\left(-\eta \sum_{h'=h+1}^{H} \sum_{(x_{h'}, a_{h'}) \in C_{h'}(x_h, a_h)} \boldsymbol{z}^t{}_{h+1:h'}(x_{h'}, a_{h'})\hat{\ell}^t_{h'}(x_{h'}, a_{h'})\right)\right] \\ =& \prod_{x_h \in C(x_{h-1}, a_{h-1})} \mathbb{E}_{\boldsymbol{z}^t, a_h}\left[\exp\left(-\eta \sum_{h'=h+1}^{H} \boldsymbol{z}^t(x_h, a_h)\right.\right. \\ &\quad \left.\left. \cdot \left(\sum_{x_{h'}, a_{h'} \in C_{h'}(x_h, a_h)} \boldsymbol{z}^t{}_{h+1:h'}(x_{h'}, a_{h'})\hat{\ell}^t_{h'}(x_{h'}, a_{h'}) + \hat{\ell}^t_h(x_h, a_h)\right)\right)\right] \\ =& \mathbb{E}_{\boldsymbol{z}^t}\left[\exp\left(-\eta \sum_{h'=h}^{H} \sum_{x_{h'}, a_{h'} \in C_{h'}(x_{h-1}, a_{h-1})} \boldsymbol{z}^t{}_{h:h'}(x_{h'}, a_{h'})\hat{\ell}^t_{h'}(x_{h'}, a_{h'})\right)\right],\end{aligned}$$

which completes the proof. □

D.2.2 Bounding the Variance of the Loss Estimate

The following lemma bounds the variance of the loss estimate.

Lemma D.4. *For and $h \in [H]$ and any $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$, it holds that $|\hat{\ell}^t_h(x_h, a_h)| \leq \frac{1}{\gamma\rho}$.*

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

$$\begin{aligned}&\left\|\boldsymbol{\phi}^{\nu^t}(x_h, a_h)\right\|_2 \\ =& \left\|- \sum_{(s_h, b_h) \in x_h \times \mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h), b_h)\boldsymbol{\phi}(s_h, a_h, b_h)\right\|_2 \\ \leq& \sum_{(s_h, b_h) \in x_h \times \mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h), b_h)\left\|\boldsymbol{\phi}(s_h, a_h, b_h)\right\|_2 \\ \overset{(i)}{\leq}& \sum_{(s_h, b_h) \in x_h \times \mathcal{B}} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h), b_h) \\ \overset{(ii)}{\leq}& 1,\end{aligned} \tag{21}$$

where (i) is due to Assumption 2.1; and (ii) follows from the proof of Lemma 2 by Kozuno et al. (2021).

Recall that $\mu^t = (1-\gamma)\hat{\mu}^t + \gamma\pi$. Let $\Phi_h^t \coloneqq \left\{\boldsymbol{\phi}^{\nu^t}(x_h, a_h)\right\}_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}$. It is then clear that

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

where (i) is because $|r_h(s_h, a_h, b_h)| \leq 1$; (ii) is by the Cauchy-Schwarz inequality; and (iii) comes from Eq. (21). □

D.2.3 Final Proof the Stability Term

We are now ready to bound the Stability term.

Lemma D.5. *The* Stability *term is bounded by*

$$\textsc{Stability} \leq \eta T X H d\,.$$

Proof. Plugging Eq. (20) into Eq. (19), we have

$$
\begin{aligned}
&\left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle + \frac{1}{\eta} \sum_{x_1 \in X_1} \log Z_1^t(x_1) \\
&= \left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle + \frac{1}{\eta} \log \Xi^t(\emptyset) \\
&= \left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle + \frac{1}{\eta} \log \mathbb{E}_{\boldsymbol{z}^t} \left[\exp \left(\underbrace{-\eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h)}_{\spadesuit} \right) \right] \\
&\overset{(i)}{\leq} \left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle + \frac{1}{\eta} \log \mathbb{E}_{\boldsymbol{z}^t} \left[1 - \eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h) + \left(\eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h) \right)^2 \right] \\
&\overset{(ii)}{\leq} \left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle - \frac{1}{\eta} \mathbb{E}_{\boldsymbol{z}^t} \left[\eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h) \right] \\
&\qquad + \frac{1}{\eta} \mathbb{E}_{\boldsymbol{z}^t} \left[\left(\eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h) \right)^2 \right] \\
&= \left\langle \hat{\mu}^t, \hat{\ell}^t \right\rangle - \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \mathbb{E}_{\boldsymbol{z}^t} \left[\boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h) \right] \\
&\qquad + \frac{1}{\eta} \mathbb{E}_{\boldsymbol{z}^t} \left[\left(\eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h) \right)^2 \right] \\
&= \frac{1}{\eta} \mathbb{E}_{\boldsymbol{z}^t} \left[\left(\eta \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \boldsymbol{z}^t{}_{1:h}(x_h, a_h) \hat{\ell}_h^t(x_h, a_h) \right)^2 \right] \\
&\overset{(iii)}{\leq} \eta \left(\sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \hat{\mu}_{1:h}^t(x_h, a_h) \right) \left(\sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \hat{\mu}_{1:h}^t(x_h, a_h) \hat{\ell}_h^t(x_h, a_h)^2 \right) \\
&\leq \eta X \left(\sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \hat{\mu}_{1:h}^t(x_h, a_h) \hat{\ell}_h^t(x_h, a_h)^2 \right), \qquad (22)
\end{aligned}
$$

where $|\spadesuit| \leq 1$ follows from setting $\gamma \geq \eta X \rho^{-1}$ and Lemma D.4 in conjunction with Assumption 3.2; (i) is from $\exp(-x) \leq 1 - x + x^2$ for $x \geq -1$; (ii) comes from $\forall x > 0, \log x \leq x - 1$; (iii) is by the Cauchy–Schwarz inequality.

The proof is then concluded by taking summation of Eq. (22) over t and using Lemma B.1.

□

E Proof of Regret Guarantees of LSFTRL

To start with, notice that $\Pi_{\max}$ is an affine subspace of $\mathbb{R}_{\geq 0}^{XA}$ 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 $\Pi_{\max} = (F + u) \cap \mathbb{R}_{\geq 0}^{XA}$ where F is a linear subspace and $u \in \Pi_{\max}$.

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

$$\Psi^\star(\boldsymbol{y}) \coloneqq \sup_{\boldsymbol{x}\in\mathbb{R}^{AX}_{\geq 0}} \langle \boldsymbol{x}, \boldsymbol{y}\rangle - \Psi(\boldsymbol{x})\,. \tag{23}$$

Also, we denote $D_{\Psi^\star}(\boldsymbol{x},\boldsymbol{y}) = \Psi^\star(\boldsymbol{x}) - \Psi^\star(\boldsymbol{y}) - \langle\nabla\Psi^\star(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle$ as the Bregman divergence induced by $\Psi^\star$. The following lemma shows the canonical regret decomposition of FTRL algorithm (Zimmert & Seldin, 2019; Lattimore & Szepesvári, 2020).

Lemma E.1. *The regret of* LSFTRL *can be decomposed as*

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

Proof. Let $\mu^\dagger\in\Pi_{\max}$ be some realization plan. For all $t\in[T]$, the instantaneous regret against $\mu^\dagger$ at step t can be decomposed into

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

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

Taking summation of the above display over t yields

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

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

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

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

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

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

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

□

The following lemma shows that the STABILITY term can be bounded by the variance of the loss estimate, which is the expected version of Lemma E.6 in Fiegel et al. (2023). We also present the proof here for completeness.

Lemma E.2. *Let $v_t=D_{\Psi^\star}(\nabla\Psi\left(\mu^t\right)-\hat{\ell}^t,\nabla\Psi\left(\mu^t\right))$ for all $t\in[T]$. Then, it holds that*

$$\textsc{Stability}=\mathbb{E}\left[\sum_{t=1}^{T}v_t\right].$$

Furthermore, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}v_t\right]\leq\mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\frac{1}{p^*_{1:h}(x_h)}\mathbb{E}^{\mu^t,\nu^t}\left[\mu^t_{1:h}(x_h,a_h)\hat{\ell}^t(x_h,a_h)^2\right]\right].$$

Proof. To begin with, for all $t\in[T]$, we define

$$f_t(u)=D_{\Psi^\star}\left(\nabla\Psi\left(\mu^t\right)-u\hat{\ell}^t,\nabla\Psi\left(\mu^t\right)\right),$$

for $u\in[0,1]$, such that $f_t(0)=0$ and $f_t(1)=v_t$. Also notice that $\operatorname{dom}(\Psi^\star)=\mathbb{R}^{XA}_{\geq 0}$ and both Ψ and $\Psi^\star$ can be decomposed according to each infoset-action pair (x_h,a_h). Specifically, we have

$$\Psi\left(\mu\right)=\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\Psi_{x_h,a_h}\left(\mu_{1:h}\left(x_h,a_h\right)\right),$$

and

$$\Psi^\star(y)=\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\Psi^\star_{x_h,a_h}\left(y\left(x_h,a_h\right)\right).$$

Then the derivative of f_t can be expressed as

$$f_t'(u)=\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\hat{\ell}^t_h\left(x_h,a_h\right)\left[\mu^t_{1:h}\left(x_h,a_h\right)-\nabla\Psi^\star_{x_h,a_h}\left(\nabla\Psi_{x_h,a_h}\left(\mu^t_{1:h}\left(x_h,a_h\right)\right)-u\hat{\ell}^t_h\left(x_h,a_h\right)\right)\right]. \tag{24}$$

Moreover, recall that we choose negative entropy as the potential function. Therefore, it holds that

$$\nabla \Psi_{x_h,a_h}\left(\mu_{1:h}\left(x_h,a_h\right)\right)=\frac{p^{\star}_{1:h}\left(x_h\right)}{\eta}\left[\log\left(p^{\star}_{1:h}\left(x_h\right)\mu_{1:h}\left(x_h,a_h\right)\right)+1\right],$$
$$\nabla \Psi^{\star}_{x_h,a_h}\left(y\left(x_h,a_h\right)\right)=\exp\left[\frac{\eta}{p^{\star}_{1:h}\left(x_h\right)}\left(y\left(x_h,a_h\right)\right)-1-\log\left(p^{\star}_{1:h}\left(x_h\right)\right)\right],$$

and

$$\begin{aligned}
&\nabla \Psi^{\star}_{x_h,a_h}\left(\nabla \Psi_{x_h,a_h}\left(\mu^t_{1:h}\left(x_h,a_h\right)\right)-u\hat{\ell}^t_h\left(x_h,a_h\right)\right)\\
&\quad=\exp\left[\frac{\eta}{p^{\star}_{1:h}\left(x_h\right)}\left(\frac{p^{\star}_{1:h}\left(x_h\right)}{\eta}\log\left(p^{\star}_{1:h}\left(x_h\right)\mu^t_{1:h}\left(x_h,a_h\right)\right)-u\hat{\ell}^t_h\left(x_h,a_h\right)\right)-\log\left(p^{\star}_{1:h}\left(x_h\right)\right)\right]\\
&\quad=\mu^t_{1:h}\left(x_h,a_h\right)\exp\left[-u\frac{\eta\hat{\ell}^t_h\left(x_h,a_h\right)}{p^{\star}_{1:h}\left(x_h\right)}\right]\\
&\quad\geq\mu^t_{1:h}\left(x_h,a_h\right)\left[1-u\frac{\eta\hat{\ell}^t_h\left(x_h,a_h\right)}{p^{\star}_{1:h}\left(x_h\right)}\right], \qquad (25)
\end{aligned}$$

where the last inequality follows from $e^{-x}\geq 1-x$ for all $x\in\mathbb{R}$.

Substituting Eq. (25) into Eq. (24) shows that

$$f'_t(u)\leq u\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\hat{\ell}^t_h\left(x_h,a_h\right)\mu^t_{1:h}\left(x_h,a_h\right)\frac{\eta\hat{\ell}^t_h\left(x_h,a_h\right)}{p^{\star}_{1:h}\left(x_h\right)}.$$

The proof is concluded by integrating the above display from 0 to 1 over u and taking the expectation on both sides. □

E.1 Additional Discussions on Assumption 4.1

Intuitively, this assumption says that the environment transition $\mathbb{P}$ and opponent's policy ν_t are *balanced* enough in the sense that $p^{\nu^t}_{1:h}\left(x_h\right)$ induced by $\mathbb{P}$ and ν^t is not too large compared with the "balanced" transition $p^{\star}_{1:h}(x_2)$ for any $x_1,x_2\in\mathcal{X}_h$ and $h\in[H]$. Indeed, consider the case where the game tree is an k-ary tree and $\mathbb{P}$ is uniform distribution at every underlying state s, then it holds that $\lambda=1$. On the other hand, the worst-case scenario is that $\lambda=\mathcal{O}(X)$ if $p^{\nu^t}_{1:H}(x_1)=1$ for some $x_1\in\mathcal{X}_H$. Nevertheless, this extreme 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. $x(s_H)=x_1$ satisfy $p_h(s_{h+1}\mid s_h,a_h,b_h)=1$ 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 ensure $\nu^t_{1:H-1}\left(y\left(s_{H-1}\right),b_{H-1}\right)=1$ by setting $\nu^t(b_h\mid y(s_h))=1$ for all (s_h,b_h) along the trajectory.

E.2 Proof of Theorem 4.2

In this section, we provide the proof of Theorem 4.2, which takes $p^{\star}$ as the transition probability function over infoset-action pairs.

Proof of Theorem 4.2. Combining Lemma E.1, E.3 and E.4, with $p^{\star}$ computed by `Computing-`$p^{\star}$, we have that

$$\begin{aligned}
\mathfrak{R}^T_{\max}&\leq \textsc{Penalty}+\textsc{Stability}\\
&\leq\frac{H}{\eta}\log\left(XA\right)+\frac{\eta}{2}THd\lambda, \qquad (26)
\end{aligned}$$

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

□

We note that leveraging β as well as Assumption 4.1 necessitates identifying a transition probability function $p^\star$ with its minimum visitation probability achieving β. Finding such $p^\star$ is done by the procedure illustrated in Appendix E.2.3.

E.2.1 Bounding the Penalty Term

The lemma below directly follows from Lemma E.5 of Fiegel et al. (2023), with its proof provided here for completeness.

Lemma E.3. *For any fixed learning rate η and transition probability function $p^\star$, it holds that*

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

Proof. It is clear that

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

where (i) comes from Lemma A.1. □

E.2.2 Bounding the Stability Term

Lemma E.4. *For any fixed learning rate η and transition probability function $p^\star$, it holds that*

$$\textsc{Stability} \leq \frac{\eta}{2} THd\lambda \, .$$

Proof. Recall that $\beta = \max_{\tilde{p} \in \mathbb{P}^\star} \min_{h \in [H], x_h \in \mathcal{X}_h} \tilde{p}_{1:h}(x_h)$. Then, one can see that

$$\begin{aligned}
\textsc{Stability} &\leq \mathbb{E}\left[\frac{\eta}{2} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{1}{p^\star_{1:h}(x_h)} \mathbb{E}^{\mu^t, \nu^t}\left[\mu^t_{1:h}(x_h,a_h) \hat{\ell}^t(x_h,a_h)^2\right]\right] \\
&\leq \mathbb{E}\left[\frac{\eta}{2} \sum_{t=1}^{T} \sum_{h=1}^{H} \frac{1}{\beta} \operatorname{tr}\left(\sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} p^{\nu^t}(x_h) \mu^t_{1:h}(x_h,a_h) \boldsymbol{\phi}^{\nu^t}(x_h,a_h) \boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top \boldsymbol{Q}^{-1}_{\mu^t,h}\right)\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\frac{\eta}{2} \sum_{t=1}^{T} \sum_{h=1}^{H} \lambda \operatorname{tr}\left(\sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu^t_{1:h}(x_h,a_h) \boldsymbol{\phi}^{\nu^t}(x_h,a_h) \boldsymbol{\phi}^{\nu^t}(x_h,a_h)^\top \boldsymbol{Q}^{-1}_{\mu^t,h}\right)\right] \\
&\overset{(ii)}{\leq} \frac{\eta}{2} THd\lambda \, ,
\end{aligned}$$

where (i) is due to Assumption 4.1; and (ii) comes from Lemma B.1. □

E.2.3 Computing $p^\star$

The procedure `Computing-`$p^\star$ can compute $p^\star$ in Eq. (9), via backward dynamic programming in $\mathcal{O}(XA)$ time.

E.3 Proof of Theorem 4.4

This section presents the proof of Theorem 4.4.

Proof of Theorem 4.4. Combining Lemma E.1, E.5 and E.6, with $p^\star_{1:h}(x_h) \equiv 1$ for all $x_h \in \mathcal{X}$, we have that

$$\begin{aligned}
\mathfrak{R}^T_{\max} &\leq \textsc{Penalty} + \textsc{Stability} \\
&\leq \frac{X(1+\log A)}{\eta} + \frac{\eta}{2} THd \, ,
\end{aligned} \tag{27}$$

which concludes the proof by noticing that $\eta = \sqrt{\frac{2X \log A}{THd}}$. □

Algorithm 5 Computing-$p^\star$

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 $C[x_h, a_h] = \sum_{x_{h+1} \in C(x_h, a_h)} f[x_{h+1}]$,
      Compute $f[x_h] = \max_{a \in \mathcal{A}} C[x_h, a]$.
    end for
  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.

E.3.1 Bounding the Penalty Term

To bound the Penalty term, we establish a refined analysis, which shaves off an $\mathcal{O}(\sqrt{A})$ factor compared with the direct combination of Lemma E.5 of Fiegel et al. (2023) and the setting of $p^\star_{1:h}(x_h) \equiv 1$.

Lemma E.5. *Setting $p^\star_{1:h}(x_h) \equiv 1$ for all $x_h \in \mathcal{X}$. For any fixed learning rate η, it holds that*

$$\text{Penalty} \leq \frac{X(1 + \log A)}{\eta}.$$

Proof.

$$
\begin{aligned}
-\Psi(\mu) &= -\frac{1}{\eta}\sum_{h=1}^{H}\Psi_h\left(p^\star_{1:h}\cdot\mu_{1:h}\right)\\
&= -\frac{1}{\eta}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu_{1:h}(x_h,a_h)\log\mu_{1:h}(x_h,a_h)\\
&= -\frac{1}{\eta}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mu_{1:h-1}(x_h)\mu_{1:h}(a_h|x_h)\left(\log\mu_{1:h-1}(x_h)+\log\mu_h(a_h|x_h)\right)\\
&= -\frac{1}{\eta}\sum_{h=1}^{H}\left(\sum_{x_h\in\mathcal{X}_h}\mu_{1:h-1}(x_h)\left(\log\mu_{1:h-1}(x_h)+\sum_{a_h\in\mathcal{A}}\mu_h(a_h|x_h)\log\mu_h(a_h|x_h)\right)\right)\\
&\le \frac{1}{\eta}\sum_{h=1}^{H}\left(\sum_{x_h\in\mathcal{X}_h}-\mu_{1:h-1}(x_h)\log\mu_{1:h-1}(x_h)+\sum_{x_h\in\mathcal{X}_h}\mu_{1:h-1}(x_h)\log A\right)\\
&\overset{(i)}{\le} \frac{1}{\eta}\sum_{h=1}^{H}\left(X_h+X_h\log A\right)\\
&= \frac{X(1+\log A)}{\eta}\,.
\end{aligned}
$$

Here (i) comes from the fact that $-x\log x\le 1$ for all $x\in[0,1]$. □

E.3.2 Bounding the Stability Term

The upper bound of Stability term when setting $p^\star_{1:h}(x_h)\equiv 1$ is guaranteed in the following lemma, the proof of which is omitted since it is essentially the same as that of Lemma E.4.

Lemma E.6. *Setting $p^\star_{1:h}(x_h)\equiv 1$ for all $x_h\in\mathcal{X}$. For any fixed learning rate η, it holds that*

$$
\begin{aligned}
\textsc{Stability} &\le \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}\mathbb{E}^{\mu^t,\nu^t}\left[\mu^t_{1:h}(x_h,a_h)\hat{\ell}^t(x_h,a_h)^2\right]\right]\\
&\le \frac{\eta}{2}THd\,.
\end{aligned}
$$

F Proof of Regret Lower Bound

In this section, we present the proof of Theorem 4.6.

Proof of Theorem 4.6. We consider an A-ary tree POMG 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. 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 POMG 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 & 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 POMG 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.

□