LEARNING IMPERFECT INFORMATION EXTENSIVE-FORM GAMES WITH LAST-ITERATE CONVERGENCE UNDER BANDIT FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

We study learning the approximate Nash equilibrium (NE) policy profile in two-player zero-sum imperfect information extensive-form games (IIEFGs) with last-iterate convergence. The algorithms in previous works studying this problem either require full-information feedback or only have asymptotic convergence rates. In contrast, we study IIEFGs in the formulation of partially observable Markov games (POMGs) with the perfect-recall assumption and bandit feedback, where the knowledge of the game is not known a priori and only the rewards of the experienced information set and action pairs are revealed to the learners in each episode. Our algorithm utilizes a negentropy regularizer weighted by a virtual transition over information set-action space. By carefully designing the virtual transition together with the leverage of the entropy regularization technique, we prove that our algorithm converges to the NE of IIEFGs with a provable finite-time convergence rate of $\widetilde{\mathcal{O}}(k^{-1/8})$ with high probability under bandit feedback, thus answering the second question of Fiegel et al. (2023) affirmatively.

1 INTRODUCTION

In imperfect information games (IIGs), players operate with limited visibility into the game's true state, necessitating strategic decision-making based on incomplete information. Notably, the concept of imperfect-information extensive-form games (IIEFGs), as introduced by Kuhn (1953), encapsulates both the intricacies of imperfect information and the sequential nature of players' moves. This framework aptly represents a broad spectrum of real-world scenarios, such as 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). Extensive research has been devoted to identifying the (approximate) Nash equilibrium (NE) (Nash Jr, 1950) within IIEFGs. Assuming the condition of perfect recall, where players possess the memory of past events and their implications, various methodologies have been employed to tackle these games. These include linear programming approaches (Koller & Megiddo, 1992; Von Stengel, 1996; Koller et al., 1996), which leverage mathematical optimization under full game knowledge, first-order optimization techniques (Hoda et al., 2010; Kroer et al., 2015; 2018; Munos et al., 2020; Lee et al., 2021; Liu et al., 2022), which iteratively refine strategies via repeated playthroughs of the games, and counterfactual regret minimization algorithms (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), which adaptively adjust strategies based on counterfactual outcomes.

In practical scenarios, IIEFGs might involve large information set and action spaces, thwarting the application of linear programming approaches for *computing* the NE in IIEFGs. In this realm, the NE in IIEFGs is typically *learned* from random samples gathered through iterative playthroughs of the game, by Monte-Carlo counterfactual regret minimization (CFR) methods (Lanctot et al., 2009; Farina et al., 2020; Farina & Sandholm, 2021) or online mirror descent (OMD) and follow-the-regularized-leader (FTRL) frameworks (Farina et al., 2021; Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023). Notably, Bai et al. (2022) devise an OMD-based approach incorporating "balanced exploration policies" to learn an ε-approximate NE with sample complexity of $\widetilde{\mathcal{O}}\left(H^3(XA+YB)/\varepsilon^2\right)$, where H is the horizon length, X, Y are the sizes of the information set space for the max- and min-player, and A and B are the sizes of the action space for the max- and min-player. This upper bound is information-theoretically optimal with respect to all parameters except H, up to logarithmic factors. Building upon Bai et al. (2022), Fiegel et al. (2023) make further strides, refining the upper bound to $\widetilde{\mathcal{O}}\left(H(XA+YB)/\varepsilon^2\right)$ by harnessing FTRL with "balanced transitions", achieving (nearly) optimal learning of IIEFGs in all parameters.

Despite the (nearly) optimal leaning of the ε-NE in IIEFGs by Bai et al. (2022); Fiegel et al. (2023), the algorithms in these works require to average all the policies generated during the running of the algorithms, so as to obtain the final policy profile with ε-NE guarantee. This is typically termed as the *average-iterate convergence*. However, in IIEFGs with large information set and action spaces, such an average operation over policy sets usually induces substantial storage and computation overhead. In cases when the policies in the games are approximated by nonlinear function approximation (*e.g.*, neural networks), which has achieved great empirical success in recent years (Moravčík et al., 2017; Brown & Sandholm, 2018), computing the averaged policy even might be not feasible due to the nonlinearity of such function approximations. This motivates the studies of the learning algorithms with the *last-iterate convergence* guarantee of games including IIEFGs (Lin et al., 2020; Wei et al., 2021a;a; Lee et al., 2021; Cai et al., 2022; Abe et al., 2023; Feng et al., 2023; Cen et al., 2023; Liu et al., 2023). Specifically, Lee et al. (2021); Liu et al. (2023) establish algorithms for learning IIEFGs with last-iterate convergence rate of $\widetilde{\mathcal{O}}(1/k)$. However, the algorithms of Lee et al. (2021); Liu et al. (2023) require full-information feedback when learning IIEFGs, and thus can not be directly applied in practical cases when the knowledge of the games is not known a priori. The above considerations naturally motivate the following question:

Can we achieve last-iterate convergence for learning IIEFGs with bandit feedback?

Indeed, the same question has also been raised by Fiegel et al. (2023). In this work, we answer this question affirmatively. The main contributions of our work are summarized as follows:

- We propose the first algorithm that learns the approximate NE of IIEFGs with provable last-iterate convergence in the bandit feedback setting. In contrast with the vanilla negentropy regularizer (Lee et al., 2021) and the dilated negentropy regularizer (Lee et al., 2021; Liu et al., 2023) used by previous works to achieve the last-iterate convergence for IIEFGs with full-information feedback, our algorithm leverages the negentropy regularizer weighted by a virtual transition over infoset-action space to regularize the game. Via constructing the loss estimator regularized by such virtual transition weighted negentropy, our algorithm avoids directly regularizing the sequence-form representation of policies and results in a desirable contraction of the KL-divergence between probability measures over the information set-action space, instead of only obtaining the KL-divergence between the sequence-form representation of policies (see Section 4.1 and Section 5.1 for details). Besides, our algorithm does not require any communication or coordination between the two players and is model-free, without requiring the knowledge of the underlying state transition probabilities and the reward functions.

- To efficiently bound the stability term in the one-step analysis of OMD with Bregman divergence induced by the virtual transition weighted negentropy regularizer, we design a virtual transition over the information set-action space that maximizes the minimum visitation probability of all the information sets (see Section 4.1 for more elaboration on this). With such a virtual transition, we finally prove that our algorithm obtains the finite-time last-iterate convergence rate for learning IIEFGs in the bandit feedback setting of $\widetilde{\mathcal{O}}((X+Y)[(XA+YB)^{1/2}+(X+Y)^{1/4}H]k^{-1/8})$ with high probability (for large enough k), where H is the horizon length, X and Y are the size of the information set spaces of the max- and min-player, A and B are the size of the action spaces of the max- and min-player, and k is the number of episodes. The methodology of our algorithm's analysis is inspired by the last-iterate convergence learning of the matrix games and the (fully observable) Markov games of Cai et al. (2023), but we provide a refined analysis specifically for IIEFGs to further sharpen the dependence on the parameters when deriving the final convergence rate (see Section 5.1 for details).

- When only obtaining the expected convergence rate is desired, our algorithm can generate a policy profile converging to the NE with a rate of $\widetilde{\mathcal{O}}((X+Y)[(X^2A+Y^2B)^{1/2}+(X+Y)^{1/4}H]k^{-1/6})$ in expectation. For the problem of learning the NE of IIEFGs in the bandit-feedback setting, we provide an $\Omega(\sqrt{XA+YB}k^{-1/2})$ lower bound of the last-iterate convergence rate.

2 Related Works

2.1 Partially Observable Markov Games (POMGs)

With perfect information, learning Markov games (MGs) can be traced back to the seminal work of Littman & Szepesvári (1996) and has since garnered extensive research attention (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). In scenarios where only imperfect information is available yet the complete knowledge of the game (transitions and rewards) is known, existing research can be categorized into three primary streams. The first stream leverages sequence-form representation of policies to recast the problem as a linear program (Koller & Megiddo, 1992; Von Stengel, 1996; Koller et al., 1996). The second stream translates the problem into a minimax optimization problem and explores first-order algorithms, as exemplified in (Hoda et al., 2010; Kroer et al., 2015; 2018; Munos et al., 2020; Lee et al., 2021; Liu et al., 2022). Lastly, the third stream addresses the problem through CFR, minimizing counterfactual regrets locally within 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).

In the realm where the knowledge of the game is either unknown or only partially accessible, the Monte-Carlo CFR algorithm introduced by Lanctot et al. (2009) pioneers the achievement of the first ε-NE result. This framework has been further generalized and extended by Farina et al. (2020); Farina & Sandholm (2021). Additionally, another line of research focuses on integrating OMD and FTRL frameworks with importance-weighted loss estimators (Farina et al., 2021; Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023). Remarkably, Bai et al. (2022) achieve an ε-approximate NE with sample complexity of $\widetilde{\mathcal{O}}\left(H^3(XA+YB)/\varepsilon^2\right)$ by employing a "balanced" dilated KL-divergence as the distance metric. Building upon this concept, Fiegel et al. (2023) utilize "balanced transitions" and attain a (nearly) optimal sample complexity of $\widetilde{\mathcal{O}}\left(H(XA+YB)/\varepsilon^2\right)$, which matches the information-theoretic lower bound up to logarithmic factors. However, we note that all the algorithms in existing works studying POMGs with bandit feedback only have *average-iterate convergence* guarantees, while we aim to establish the *last-iterate convergence* guarantee.

2.2 Last-iterate Convergence Learning in Games

With full-information feedback, learning in games with last-iterate convergence guarantee has been investigated in strongly monotone games (Mokhtari et al., 2020; Jordan et al., 2024), monotone games (Golowich et al., 2020; Cai et al., 2022; Gorbunov et al., 2022; Cai & Zheng, 2023), Markov games (Cen et al., 2021; 2023), and IIEFGs (Lee et al., 2021; Liu et al., 2023; Bernasconi et al., 2024).

Recently, motivated by the fact that it might be restrictive to require full knowledge of the (noisy) gradient as in the full-information feedback setting, a growing body of works has studied learning in games with last-iterate convergence guarantee in the bandit feedback setting including strongly monotone games (Bravo et al., 2018; Lin et al., 2021) (Bravo et al., 2018; Hsieh et al., 2019; Lin et al., 2021; Drusvyatskiy et al., 2022; Huang & Hu, 2023), matrix games (Cai et al., 2023) and Markov games (Wei et al., 2021b; Chen et al., 2022; Cai et al., 2023). However, the algorithm of Wei et al. (2021b) needs coordinated updates and some prior knowledge of the game, and the algorithm of Chen et al. (2022) requires the players to inform the opponent about the entropy of their own policies. Amongst these works, Cai et al. (2023) remove all the coupling requirements, achieving last-iterate convergences of $\widetilde{\mathcal{O}}(k^{-1/8})$ for matrix games and of $\widetilde{\mathcal{O}}(k^{-1/9+\varepsilon})$ for any $\varepsilon > 0$ for irreducible Markov games. We note that all existing works study fully-observable Markov games, while we aim to establish uncoupled algorithms for learning IIEFGs in the formulation of partially-observable Makov games, without requiring the knowledge of the games.

3 Preliminaries

For ease of exposition, we consider IIEFGs in the formulation of POMGs and introduce the preliminaries of them in this section, following previous works (Kozuno et al., 2021; Bai et al., 2022).

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

- H is the horizon length;
- $\mathcal{S} = \bigcup_{h\in[H]} \mathcal{S}_h$ is the finite state space with $\mathcal{S}_h$ as the state space at step h. $S = \sum_{h=1}^{H} S_h$ is the size of $\mathcal{S}$ where $|\mathcal{S}_h| = S_h, \forall h \in [H]$;
- $\mathcal{X} = \bigcup_{h\in[H]} \mathcal{X}_h$ is the finite space of information sets (short for *infosets* in the following) for the max-player, where $\mathcal{X}_h = \{x(s) : s \in \mathcal{S}_h\}$ is the set of the infosets at step h with $x : \mathcal{S} \to \mathcal{X}$ as the emission function. $X = \sum_{h=1}^{H} X_h$ is the size of $\mathcal{X}$ with $|\mathcal{X}_h| = X_h$. The finite space of infosets $\mathcal{Y} = \bigcup_{h\in[H]} \mathcal{Y}_h$ for the min-player and its size are defined analogously;
- $\mathcal{A}$ with $|\mathcal{A}| = A$ and $\mathcal{B}$ with $|\mathcal{B}| = B$ are the finite action spaces for the max-player and min-player, respectively;
- $\mathbb{P} = \{p_0(\cdot) \in \Delta_{\mathcal{S}_1}\} \bigcup \{p_h(\cdot|s_h, a_h, b_h) \in \Delta_{\mathcal{S}_{h+1}}\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B},h\in[H-1]}$ are the state transition probabilities, where $p_0(\cdot)$ is the probability distribution of initial states, $p_h(s_{h+1}|s_h, a_h, b_h)$ is the probability of transitioning to the next state s_{h+1} conditioned on (s_h, a_h, b_h) at step h, and $\Delta_{\mathcal{S}_h}$ denotes the probability simplex over $\mathcal{S}_h$;
- $r = \{r_h(s_h, a_h, b_h) \in [0, 1]\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B},h\in[H]}$ are the (randomized) reward functions with $\bar{r}_h(s_h, a_h, b_h)$ as mean for each $r_h(s_h, a_h, b_h)$.

Learning Protocol Define the max-player's stochastic policy as $\mu = \{\mu_h\}_{h\in[H]}$, where $\mu_h^k : \mathcal{X}_h \to \Delta_{\mathcal{A}}$ denotes the policy at step h during episode k. The set of all such policies for the max-player is denoted by $\Pi_{\max}$. Analogously, the min-player's stochastic policy is specified as $\nu = \{\nu_h\}_{h\in[H]}$, with $\nu_h^k : \mathcal{Y}_h \to \Delta_{\mathcal{B}}$ being the policy at step h during episode k, and the set of all min-player policies is denoted by $\Pi_{\min}$. The game proceeds in a finite number of episodes. At the commencement of episode k, the max-player selects a stochastic policy $\mu^k \in \Pi_{\max}$, while the min-player chooses $\nu^k \in \Pi_{\min}$. Meanwhile, an initial state s_1^k is sampled from the distribution $p_0(\cdot)$ by the environment. During each step h within an episode, the max-player and min-player observe their respective infosets $x_h^k := x(s_h^k)$ and $y_h^k := y(s_h^k)$, but they do not directly observe the underlying state s_h^k. Given x_h^k, the max-player samples and executes an action $a_h^k \sim \mu_h^k(\cdot|x_h^k)$, while the min-player concurrently takes an action $b_h^k \sim \nu_h^k(\cdot|y_h^k)$. Upon taking these actions, the max-player and min-player receive rewards $r_h^k := r_h(s_h^k, a_h^k, b_h^k)$ and $-r_h^k$, respectively. Subsequently, the game transitions to the next state $s_{h+1}^t \sim p_h(\cdot|s_h^k, a_h^k, b_h^k)$. The k-th episode will terminate after actions a_H^k and b_H^k are taken conditioned on x_H^k and y_H^k.

Perfect Recall and Tree Structure Following prior works (Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023), we assume that the POMGs adhere to the *tree structure* and the *perfect recall* condition, as defined by Kuhn (1953). Explicitly, the tree structure signifies that for any step $h = 2, \ldots, H$ and state $s_h \in \mathcal{S}_h$, there exists a *unique* path $(s_1, a_1, b_1, \ldots, s_{h-1}, a_{h-1}, b_{h-1})$ culminating in s_h. The perfect recall condition, meanwhile, is fulfilled for both players, implying that for any $h = 2, \ldots, H$ and any infoset $x_h \in \mathcal{X}_h$ of the max-player (analogously for the min-player), there exists a *unique* history $(x_1, a_1, \ldots, x_{h-1}, a_{h-1})$ leading to x_h. Furthermore, we introduce the notation $C_{h'}(x_h, a_h) \subset \mathcal{X}_{h'}$ to represent the set of descendants of the infoset-action pair (x_h, a_h) at step $h' \geqslant h$. Also, we define $C_{h'}(x_h) := \bigcup_{a_h\in\mathcal{A}} C_{h'}(x_h, a_h)$ as the union of descendants across all actions at x_h, and for convenience, let $C(x_h, a_h) := C_{h+1}(x_h, a_h)$ signify the immediate descendants at the subsequent step.

Sequence-form Representations For any pair of product policies (μ, ν), the tree structure and the perfect recall condition facilitate the *sequence-form representation* of the reaching probability for the state-action tuple (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 $p_{1:h}(s_h) = p_0(s_1) \prod_{h'=1}^{h-1} p_{h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'})$ denotes the sequence-form transition probability, and $\mu_{1:h}(x_h, a_h) := \prod_{h'=1}^{h} \mu_{h'}(a_{h'}|x_{h'})$ and $\nu_{1:h}(y_h, b_h) := \prod_{h'=1}^{h} \nu_{h'}(b_{h'}|y_{h'})$ represent the sequence-form policies of the max- and min player, respectively. Under the sequence-form representation, we adopt a slight abuse of notation for μ and ν by interpreting them as $\mu = \{\mu_{1:h}\}_{h\in[H]}$

and $\nu = \{\nu_{1:h}\}_{h\in[H]}$.[1] Furthermore, it is clear that $\Pi_{\max}$ constitutes a convex compact subspace of $\mathbb{R}^{XA}$ that adheres to the constraints $\mu_{1:h}(x_h, a_h) \geqslant 0$ and $\sum_{a_h\in\mathcal{A}} \mu_{1:h}(x_h, a_h) = \mu_{1:h-1}(x_{h-1}, a_{h-1})$, where (x_{h-1}, a_{h-1}) is such that $x_h \in C(x_{h-1}, a_{h-1})$ (with the understanding that $\mu_{1:0}(x_0, a_0) = 1$ as a base case).

Learning Objective In this work, we consider the learning objective of finding an approximate NE of the POMG. Specifically, for any $\varepsilon \geqslant 0$, an ε-approximate NE is a pair of product policy (μ, ν) satisfying

$$\text{NEGap}(\mu, \nu) \coloneqq \max_{\mu^\dagger\in\Pi_{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger\in\Pi_{\min}} V^{\mu,\nu^\dagger} \leqslant \varepsilon\,, \tag{2}$$

where $V^{\mu,\nu} = \mathbb{E}_{\mu,\nu}\left[\sum_{h=1}^{H} r_h(s_h, a_h, b_h)\right]$ the value function of (μ, ν) with the expectation taken over the randomness of the product policy pair (μ, ν) and the environment. It is known that using regret to NE conversion, an approximate NE can be obtained by averaging all the policies $\{\mu\}_{k=1}^{K}$ of the max-player generated by an algorithm with sublinear regret (similarly for the min-player) to obtain the average policy pair $(\bar{\mu}, \bar{\nu})$ (see, *e.g.*, Theorem 1 of Kozuno et al. (2021)). This is the so-called *average-iterate convergence* of learning NE. In this work, we are interested in finding the ε-NE with the (finite-time) *last-iterate convergence* guarantee; that is, the algorithm is required to generate an approximate NE policy profile (μ^k, ν^k) such that $\text{NEGap}(\mu^k, \nu^k) \leqslant \varepsilon_k$ in each episode for finite-time k.

Information Available to the Players In this work, we consider learning POMGs in the bandit feedback setting, where in each episode k, the max-player only observes her experienced trajectory $(x_1^k, a_1^k, r_1^k, \ldots, x_H^k, a_H^k, r_H^k)$ of infosets, actions, and rewards, but not the underlying states or the opponent's infosets and actions. Additionally, the max-player does not have knowledge about the policies adopted by the min-player and also can not receive any information from the min-player and vice versa. Besides, there is no shared randomness between both players; that is, the algorithms of both players need to be fully uncoupled from each other.

Additional Notations We slightly abuse the notation to view x_h as the set $\{s \in \mathcal{S}_h : x(s) = x_h\}$, when writing $s \in x_h$. Given 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 define $\langle \mu, f\rangle \coloneqq \sum_{h\in[H],(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h, a_h) f_h(x_h, a_h)$. We denote by $\mathcal{F}^k$ the σ-algebra generated by the random variables $\{(s_h^t, a_h^t, b_h^t, r_h^t)\}_{h\in[H],t\in[k]}$. For brevity, we abbreviate the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}^k]$ as $\mathbb{E}^k[\cdot]$. Throughout this paper, the notation $\widetilde{\mathcal{O}}(\cdot)$ suppresses all logarithmic factors.

4 Algorithm

In this section, we introduce the proposed algorithm, detailed in Algorithm 1.

4.1 From Sequence-form Representations to Probability Measures over Infoset-Action Space

With sequence-form representations, we first reformulate the IIEFG into the following bilinear game:

$$f(\mu, \nu) = \mu^\top \boldsymbol{G} \nu\,, \tag{4}$$

where $\boldsymbol{G} \in \mathbb{R}^{XA\times YB}$ is the loss matrix with $\boldsymbol{G}[(x_h, a_h), (y_h, b_h)] = \sum_{s_h\in x_h\cap y_h} p_{1:h}(s_h)\,(1 - r_h(s_h, a_h, b_h))$. In this manner, the learning objective is equivalent to finding (μ, ν) such that $\text{NEGap}(\mu, \nu) = \sup_{\mu^\dagger\in\Pi_{\max},\nu^\dagger\in\Pi_{\min}} f(\mu, \nu^\dagger) - f(\mu^\dagger, \nu) \leqslant \varepsilon$. At a high level, we apply the entropy regularizing technique to perturb the bilinear form of the game, as defined in Eq. (4), into a strongly convex-strongly concave structure, ensuring convergence to both the NE of the perturbed game (and thus the NE of the original game in Eq. (4)). This approach

[1] The sequence-form representation of policies is defined in a top-down manner and is equivalent to the "treeplex" space of policies defined in a bottom-up manner (see, *e.g.*, Lee et al. (2021)).

Algorithm 1 OMD with Virtual Transition Weighted Negentropy Regularization (max-player)

Input: $\eta_k = k^{-\alpha_\eta}$, $\gamma_k = k^{-\alpha_\gamma}$, $\varepsilon_k = k^{-\alpha_\varepsilon}$.
Initialize: $\mu_1(a_h|x_h) = \frac{1}{A}$, $\forall (x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}, \forall h \in [H]$. Set p^x computed by Algorithm 2.
for $k = 1, \cdots,$ **do**
  for $h = 1, \cdots, H$ **do**
    Observes x_h^k, executes $a_h^k \sim \mu_h^k(\cdot|x_h^k)$ and receives r_h^k.
    For all $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$, set entropy regularized loss estimator as

$$\widehat{\ell}_h^k(x_h, a_h) = \frac{\mathbb{I}_h^k\{x_h, a_h\}}{\mu_{1:h}^k(x_h, a_h) + \gamma_k}(1 - r_h^k) + \varepsilon_k \cdot p_{1:h}^x(x_n) \log[p_{1:h}^x \cdot \mu_{1:h}^k](x_h, a_h)\,.$$

  end for
  Update policy

$$\mu^{k+1} = \underset{\mu \in \Pi_{\max}^{k+1}}{\arg\min}\, \eta_k \langle \mu, \widehat{\ell}^k \rangle + D_\psi(\mu, \mu^k)\,, \tag{3}$$

  where $\Pi_{\max}^{k+1} = \{\mu \in \Pi_{\max} : \mu(a_h|x_h) \geqslant \frac{1}{A(k+1)}, \forall (x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}, \forall h \in [H]\}$.
end for

Algorithm 2 Computing virtual transition p^x (max-player)

Input: Game tree structure of $\mathcal{X} \times \mathcal{A}$.
Initialization: Sequence-form representation of virtual transition $q \in \mathbb{R}^X$; array of maximized number of descendant infoset $c \in \mathbb{R}^X$, $d \in \mathbb{R}^{XA}$. For all x_H in $\mathcal{X}_H$, set $c[x_H] = 1$.
for $h = H - 1$ to 1 **do**
  for x_h in $\mathcal{X}_h$ **do**
    for a_h in $\mathcal{A}$ **do**
      Compute $d[x_h, a_h] = \sum_{x_{h+1} \in C(x_h, a_h)} c[x_{h+1}]$.
    end for
    Compute $c[x_h] = \max_{a \in \mathcal{A}} d[x_h, a]$.
  end for
end for
for x_1 in $\mathcal{X}_1$ **do**
  Compute $q_{1:1}(x_1) = \frac{c[x_1]}{\sum_{x_1 \in \mathcal{X}_1} c[x_1]}$.
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 $q_{1:h+1}(x_{h+1}) = q_{1:h}(x_h) \cdot \frac{c[x_{h+1}]}{\sum_{x_{h+1} \in C(x_h, a_h)} c[x_{h+1}]}$.
    end for
  end for
end for
return q.

builds upon previous research that has explored last-iterate convergence learning in Markov games with full-information feedback (Cen et al., 2021; Chen et al., 2022; Cen et al., 2023), matrix games and Markov games with bandit feedback (Cai et al., 2023), and IIEFGs with full-information feedback (Liu et al., 2023). Specifically, we consider the following perturbed game as a surrogate:

$$f_k(\mu, \nu) = \mu^\top \boldsymbol{G} \nu + \varepsilon_k \psi(\mu) - \varepsilon_k \psi(\nu)\,, \tag{5}$$

where ψ is some strongly convex regularizer to be used in OMD and $\varepsilon_k > 0$ serves as the knob to control the strength of the entropy regularization in episode k. Intuitively, due to the strongly convex-strongly concave property of the perturbed game, one is able to find the approximate NE of it with last-iterate convergence using OMD. On the other hand, by gradually decreasing ε_k to be moderately small, the approximate NE of the perturbed game in Eq. (5) will also serve as an approximate NE of the original game in Eq. (4).

The crucial aspect lies in selecting an appropriate regularizer ψ. Initially, the first candidate that might come to mind is the utilization of the vanilla negentropy regularizer $\psi(\mu) = \sum_{h,x_h,a_h} \mu_{1:h}(x_h, a_h) \log \mu_{1:h}(x_h, a_h)$, which has been utilized to achieve the last-iterate convergence for IIEFGs with full-information feedback (Lee et al., 2021) and matrix games, the special case of IIEFGs, with bandit feedback (Cai et al., 2023). However, in IIEFGs with bandit feedback, though using the vanilla negentropy regularizer results in a convergence of the Bregman divergence, it is generally hard to control the NE gap since it directly regularizes the sequence-form representation policies. The other natural approach is considering using the dilated negentropy $\psi(\mu) = \sum_{h,x_h,a_h} \mu_{1:h}\left(x_h, a_h\right) \log\left(\frac{\mu_{1:h}(x_h,a_h)}{\mu_{1:h}(x_h)}\right)$ (Kroer et al., 2015; Kozuno et al., 2021). Indeed, the dilated negentropy has also been used to achieve the last-iterate convergence of the IIEFGs with full-information feedback (Lee et al., 2021; Liu et al., 2023; Bernasconi et al., 2024). However, in contrast with the full-information feedback setting, leveraging the entropy regularization technique to obtain the finite-time convergence guarantee in the bandit feedback setting requires the probability of selecting each action a_h given each infoset x_h being lower bounded to prevent the stability term in the analysis of OMD from being prohibitively largely. This essentially requires constraining the optimization of OMD onto a subset of the entire space of the sequence-form representations of policies $\Pi_{\max}$. Nevertheless, this will also make the stability term of OMD using the dilated negentropy in conjunction with the regularization technique hard to control, as bounding the stability term of the OMD with dilated negentropy critically relies upon its closed-form update solution (see, *e.g.*, Lemma 7 of Kozuno et al. (2021)), which no longer holds in the case where the policy update of OMD is constrained onto a subset of $\Pi_{\max}$.

To cope with the above difficulties, we instead consider using the negentropy regularizer weighted by a kind of *virtual transition* p^x over the infoset-action space $\mathcal{X} \times \mathcal{A}$:

$$\psi_{p^x}(\mu) = \sum_{h,x_h,a_h} p^x_{1:h}(x_h)\mu_{1:h}(x_h, a_h) \log\left(p^x_{1:h}(x_h)\mu_{1:h}(x_h, a_h)\right) ,$$

where $p^x_h(\cdot|x_h, a_h) \in \Delta_{C(x_h,a_h)}$ is a transition probability over $\mathcal{X}_h \times \mathcal{A} \times \mathcal{X}_{h+1}$ and $p^x_{1:h}(x_h) = p^x_0(x_1) \prod_{h'=1}^{h-1} p^x_{h'}(x_{h'+1}|x_{h'}, a_{h'})$ is its sequence-form representation. Note that $p^x_h(x_{h+1}|x(s_h), a_h)$ is not necessarily to be the true transition probability $\mathbb{P}^{\mu^k,\nu^k}(x_{h+1}|x(s_h), a_h) = \sum_{s_{h+1}\in x_{h+1}, b_h\in\mathcal{B}} p(s_{h+1}|s_h, a_h, b_h)\nu^k(b_h|y(s_h))$ experienced by the max-player in episode k. Also, notice that $\psi_{p^x}(\cdot)$ is dependent on the chosen virtual transition p^x and we drop the dependence in the subscript of $\psi_{p^x}(\cdot)$ on p^x when the context is clear for brevity. We remark that similar ideas leveraging negentropy weighted by the transition over infoset-action space have also been exploited by Bai et al. (2022); Fiegel et al. (2023). However, we would like to underscore that the design of our virtual transition $p^\star$ over infoset-action space is different from those of Bai et al. (2022); Fiegel et al. (2023) and we aim to establish the last-iterate convergence of IIEFGs while they can only guarantee the average-iterate convergence, necessitating different theoretical analysis. Besides, one can see that the constructed virtual transition p^x is well-defined by the perfect recall condition and $p^x_{1:h} \cdot \mu_{1:h}$ with $[p^x_{1:h} \cdot \mu_{1:h}](x_h, a_h) = p^x_{1:h}(x_h)\mu_{1:h}(x_h, a_h)$ is a probability measure over the infoset-action space $\mathcal{X}_h \times \mathcal{A}$ at step h. Therefore, we actually regularize the probability measures over $\mathcal{X}_h \times \mathcal{A}$ instead of directly regularizing the sequence-form representation μ, which tackles the difficulties of using the vanilla negentropy and the dilated negentropy as mentioned above. The other nice property of virtual transition weighted negentropy is that $D_\psi(\mu_1, \mu_2) = \mathrm{KL}(p^x\mu_1, p^x\mu_2)$, facilitating bounding the final NE gap as we shall see in Section 5.1.

With regularizer ψ specified, the derivative of $f_k(\mu, \nu)$ w.r.t. $\mu(x_h, a_h)$ is $\frac{\partial f_k(\mu,\nu)}{\partial \mu_{1:h}(x_h,a_h)} = \boldsymbol{G}\nu\left[(x_h, a_h)\right] + \varepsilon_k \cdot p^x_{1:h}(x_n)\left[\log[p^x_{1:h} \cdot \mu_{1:h}](x_h, a_h) + 1\right]$. Since $[p^x_{1:h} \cdot \mu_{1:h}] \in \Delta_{\mathcal{X}_h\times\mathcal{A}}$ for any μ, the constant 1 in the above display does not affect the optimization of OMD. On the other hand, in the bandit feedback setting, an (optimistically biased) loss estimate $\frac{\mathbb{I}\{x_h,a_h\}}{\mu^k_{1:h}(x_h,a_h)+\gamma_k}(1 - r^k_h)$ of $\boldsymbol{G}\nu\left[(x_h, a_h)\right]$ in episode k is constructed (Kozuno et al., 2021), where $\gamma_k > 0$ is the implicit exploration parameter (Neu, 2015). This specifies the final entropy regularized loss estimator used by Algorithm 1 on Line 6.

With the constructed loss estimator, Algorithm 1 then uses OMD to update policy. Since now the entropy regularized loss estimator is considered, the variance of the loss estimator will be prohibitively large if running OMD on the entire space of the sequence-form representations $\Pi_{\max}$, eventually leading to an unbounded stability term of OMD. Hence we constrain the feasible set of the OMD as a subset $\Pi_{\max}^{k+1}$ of $\Pi_{\max}$, where each $\mu \in \Pi_{\max}^{k+1}$ satisfying $\mu(a_h|x_h)$ is lower bounded for all $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$ and $h \in [H]$ (Line 8).

4.2 Virtual Transition with Maximized Minimum Visitation Probability

As elaborated in Section 4.1, our Algorithm 1 leverages a virtual transition weighted negentropy to regularize the loss estimator and induce the Bregman divergence used in OMD. It remains to specify an appropriate virtual transition p^x. The upside of employing such virtual transition p^x lies in that it implicitly helps to operate the update of OMD in the space of probability measures over infoset-action pairs instead of the sequence-form representations of policies. However, this also comes at the expense of enlarging the stability term of OMD. Specifically, upon applying the virtual transition to weight the negentropy, the stability term associated with OMD at each information set x_h will be enlarged by (approximately) a multiplicative factor of $1/p^x(x_h)$. This enlargement arises intuitively from the fact that, at each x_h, the Bregman divergence induced by ψ undergoes a downscaling, proportional to $p^x(x_h)$, thereby resulting in a relative increase in the stability term. Therefore, to ensure that the stability term is well-controlled, we design the following p^x which maximizes the minimum visitation probability of all x_h in its sequence-form representation:

$$p^x = \arg\max_{q \in \mathbb{P}^x} \min_{x_h \in \mathcal{X}_h, h \in [H]} q_{1:h}(x_h) \,. \tag{6}$$

In the above display, we denote by $\mathbb{P}^x$ the set of all the valid virtual transitions over infoset-action space. We note that such a virtual transition p^x can be efficiently computed by Algorithm 2 via backward dynamic programming.

Computation Due to the fact the update of OMD is now constrained onto a subset $\Pi_{\max}^k$ of the entire space $\Pi_{\max}$ of the sequence-form representation policies, the computation of Eq. (3) generally does not have a closed-form solution. We hereby provide an algorithm, which computes an approximate solution to Eq. (3), detailed in Algorithm 3. In particular, Algorithm 3 utilizes a Frank–Wolfe-type procedure to compute the update in Eq. (3). In particular, there will be T iterations in Algorithm 3, and in each iteration t, the policy will be updated towards the direction that minimizes the gradient of the objective function w.r.t. policy $\mu^{(t-1)}$ in iteration $t-1$ by dynamic programming in Algorithm 4. We defer the details of Algorithm 3 and Algorithm 4 to Appendix F.

5 Analysis

In this section, we first present the upper bound of the last-iterate convergence rate of our Algorithm 1. Then the lower bound for the problem of learning IIEFGs with bandit feedback and last-iterate convergence guarantee will be provided.

5.1 Upper Bound of Last-iterate Convergence

Theorem 5.1. *If Algorithm 1 is adopted by both players, for any $k \geqslant 1$, with probability at least $1 - \widetilde{\mathcal{O}}(\delta)$, it holds that*

$$\mathrm{NEGap}(\mu^k, \nu^k) = \mathcal{O}\Big(\Big[(XA+YB)^{\frac{1}{2}} k^{-\frac{1}{8}} + (XA+YB)^{\frac{1}{2}} Hk^{-\frac{3}{8}} + \big(X^2A+Y^2B\big)^{\frac{1}{2}} k^{-\frac{1}{4}} + (X+Y)^{\frac{1}{4}} Hk^{-\frac{1}{8}}\Big]$$
$$\cdot (X+Y)\left(\log\left(XAk/\delta\right) + \log\left(YBk/\delta\right)\right) \log^{\frac{1}{2}}(k) + k^{-\frac{1}{8}} H(\ln(XA) + \ln(YB)) + (XAB + YBH)/k\Big) \,.$$

Remark 5.2. *Ignoring the poly-logarithmic terms and when k is large enough (specifically, $k \geqslant \max\{H^4, {(X^2A+Y^2B)^4}/{(XA+YB)^4}, {(XA+YB)^{8/7}}/{(X+Y)^{10/7}}\}$), we have $\mathrm{NEGap}(\mu^k, \nu^k) = \widetilde{\mathcal{O}}((X+Y)[(XA+YB)^{1/2} + (X+Y)^{1/4}H]k^{-1/8})$. Besides, when only obtaining an expected last-iterate convergence rate is desired, our Algorithm 1 has an improved last-iterate convergence rate of $\widetilde{\mathcal{O}}((X+Y)[(X^2A+Y^2B)^{1/2} + (X+Y)^{1/4}H]k^{-1/6})$ in expectation, the details of which are deferred to Appendix C. Though the last-iterate convergence rate of our Algorithm 1 is inferior to the $\widetilde{\mathcal{O}}(1/k)$ convergence rate by Lee et al. (2021); Liu et al. (2023), we note that both their algorithms can only work in the full-information setting. Further, we remark that the algorithm of Lee et al. (2021) needs the assumption that the NE of the IIEFG considered is unique, and the algorithm of Liu et al. (2023) requires both players being controlled by a central controller, and thus the algorithm* *of Liu et al. (2023) is not uncoupled. In contrast, our algorithm can work in the bandit feedback setting, is fully uncoupled between the two players, and can still guarantee a regret of order $\widetilde{\mathcal{O}}(k^{7/8})$ when the opponent of the max-player is an adversary. More importantly, we show in Section 5.2 that the lower bound of the convergence rate for learning IIEFGs with bandit feedback, last-iterate convergence guarantee, and uncoupled algorithms will be of order $\Omega(k^{-1/2})$ (for large enough k).*

Proof Sketch of Theorem 5.1 We postpone the complete proof of Theorem 5.1 to Appendix B. Here we provide a proof sketch of it.

We denote by $\xi^{k,\star} := (\mu^{k,\star}, \nu^{k,\star})$ the unique NE in the regularized game f_k in Eq. (5), where there is only a unique NE since f_k is strongly convex in μ and strongly concave in ν. We first show that in each episode k, the product policy $\xi^k := (\mu^k, \nu^k)$ generated by the algorithm will approach $\xi^{k,\star}$ close enough by showing that the Bregman divergence $D_\psi(\xi^{k,\star}, \xi^k)$ is an (approximate) contraction mapping. In particular, we show that

$$\begin{aligned} D_\psi\left(\xi^{k+1,\star}, \xi^{k+1}\right) \lesssim & \left(1-\eta_k \varepsilon_k\right) D_\psi\left(\xi^{k,\star}, \xi^k\right) + \eta_k^2\left(X \underline{\tau}_k + Y \bar{\tau}_k\right) + \eta_k^2\left(X^2 A + Y^2 B\right) \\ & + \eta_k \rho_k + \eta_k \sigma_k + \eta_k^2 \varepsilon_k^2 H^2 (XA + YB) + \omega_k \,, \end{aligned} \tag{7}$$

where we denote

$$\begin{aligned}
\underline{\tau}_k &= \frac{1}{X} \sum_{h, x_h, a_h} \frac{1}{p_{1:h}^x(x_h)} \left(\frac{\mathbb{I}_h^k\{x_h, a_h\}}{\mu_{1:h}^k(x_h, a_h) + \gamma_k} - 1 \right), \\
\bar{\tau}_k &= \frac{1}{Y} \sum_{h, y_h, b_h} \frac{1}{p_{1:h}^y(y_h)} \left(\frac{\mathbb{I}_h^k\{y_h, b_h\}}{\nu_{1:h}^k(y_h, b_h) + \gamma_k} - 1 \right), \\
\rho_k &= \sum_{h, x_h, a_h} \mu_{1:h}^k(x_h, a_h) \left[\left(\boldsymbol{G}\nu^k\right)[(x_h, a_h)] - \left(\frac{\mathbb{I}_h^k\{x_h, a_h\}}{\mu_{1:h}^k(x_h, a_h) + \gamma_k} \left(1 - r_h^k\right) \right) \right] \\
&\quad + \sum_{h, y_h, b_h} \nu_{1:h}^k(y_h, b_h) \left[\left(1 - \left(\boldsymbol{G}^\top \mu^k\right)[(y_h, b_h)]\right) - \frac{\mathbb{I}_h^k\{y_h, b_h\}}{\nu_{1:h}^k(y_h, b_h) + \gamma_k} r_h^k \right], \\
\sigma_k &= \sum_{h, x_h, a_h} \mu_{1:k}^{k,\star}(x_h, a_h) \left[\left(\boldsymbol{G}\nu^k\right)[(x_h, a_h)] - \left(\frac{\mathbb{I}_h^k\{x_h, a_h\}}{\mu_{1:h}^k(x_h, a_h) + \gamma} \left(1 - r_h^k\right) \right) \right] \\
&\quad + \sum_{h, y_h, b_h} \nu_{1:h}^{k,\star}(y_h, b_h) \left[\frac{\mathbb{I}_h^k\{y_h, b_h\}}{\nu_{1:h}^k(y_h, b_h) + \gamma_k} r_h^k - \left(1 - \left(\boldsymbol{G}^\top \mu^k\right)[(y_h, b_h)]\right) \right], \\
\omega_k &= D_\psi\left(\mu^{k+1,\star}, \mu^{k+1}\right) - D_\psi\left(\mu^{k,\star}, \mu^{k+1}\right) + D_\psi\left(\nu^{k+1,\star}, \nu^{k+1}\right) - D_\psi\left(\nu^{k,\star}, \nu^{k+1}\right).
\end{aligned}$$

Expanding the above recursion, we can bound $D_\psi\left(\xi^{k+1,\star}, \xi^{k+1}\right)$ as

$$\begin{aligned} D_\psi\left(\xi^{k+1,\star}, \xi^{k+1}\right) \lesssim & \underbrace{\sum_{i=1}^k w_k^i \eta_i \rho_i}_{\textbf{Term 1}} + \underbrace{\sum_{i=1}^k w_k^i \eta_i \sigma_i}_{\textbf{Term 2}} + \underbrace{(XA + YB) H^2 \sum_{i=1}^k w_k^i \left(\eta_i \varepsilon_i\right)^2}_{\textbf{Term 3}} \\ & + \underbrace{\sum_{i=1}^k w_k^i \eta_i^2 \left(X \underline{\tau}_i + Y \bar{\tau}_i\right)}_{\textbf{Term 4}} + \underbrace{\sum_{i=1}^k w_k^i \eta_i^2 \left(X^2 A + Y^2 B\right)}_{\textbf{Term 5}} + \underbrace{\sum_{i=1}^k w_k^i \omega_i}_{\textbf{Term 6}}, \end{aligned} \tag{8}$$

where $w_k^i = \prod_{j=i+1}^k (1 - \eta_j \varepsilon_j)$ is the contraction parameter. Then we bound each of the above terms in by Lemma B.4 - Lemma B.9 in Appendix B.2. Note that we follow a similar analysis scheme of Cai et al. (2023) to bound the last-iterate convergence of learning matrix games with bandit feedback. However, we also remark that the straightforward application of their analysis will not address our problem of learning IIEFGs with bandit feedback, since we leverage a different regularizer and a new virtual transition p^x computed by Algorithm 2, which serves as a core ingredient of the analysis in deriving the contraction of Eq. (7) and bounding **Term 6**. Besides, compared with the analysis of Cai et al. (2023), the additional **Term 5** in Eq. (8) comes from the fact that we establish a refined analysis in the case of IIEFGs to further sharpen the dependence on X and A (as well as Y and B) of the final convergence rate.

Further, one can see that the NE policy profile $\xi^{k,\star}$ of the perturbed game in Eq. (5) is also an approximate NE of the original game in Eq. (4), enabling to bound $\mathrm{NEGap}(\xi^k)$ using $\mathrm{NEGap}(\xi^{k,\star})$ together with the distance between ξ^k and $\xi^{k,\star}$ weighted by the virtual transitions as bellow:

$$\mathrm{NEGap}(\xi^k) \leqslant \mathrm{NEGap}(\xi^{k,\star}) + X \left\|p^x \left(\mu^k - \mu^{k,\star}\right)\right\|_1 + Y \left\|p^y \left(\nu^k - \nu^{k,\star}\right)\right\|_1 , \tag{9}$$

where $\mathrm{NEGap}(\xi^{k,\star})$ can be controlled by Lemma B.2. Due to the constructed virtual transition p^x and p^y, the second and the third term in Eq. (9) are actually the ℓ_1-norm of the difference between the probability measures over infoset-action spaces, which thus turns out to be bounded by $\mathcal{O}(\sqrt{\mathrm{KL}\left(p^x\mu^{k,\star}, p^x\mu^k\right)})$ and $\mathcal{O}(\sqrt{\mathrm{KL}\left(p^y\nu^{k,\star}, p^y\nu^k\right)})$ by Pinsker's inequality. Also, thanks to the virtual transition weighted negentropy ψ, one can see that $\mathrm{KL}\left(p^x\mu^{k,\star}, p^x\mu^k\right) = D_\psi(\mu^{k,\star}, \mu^k)$ (and similarly on the min-player side). Therefore, the proof can be concluded by substituting Eq. (8) into Eq. (9) and then using Lemma B.2 and Lemma B.4 - Lemma B.9.

5.2 Lower Bound of Last-iterate Convergence

Theorem 5.3. *For any algorithm* Alg *that both players adopt to generate policy profile* (μ^k, ν^k) *and is uncoupled between both players, there exists an IIEFG instance such that the lower bound of the last-iterate convergence of learning this IIEFG in the bandit-feedback setting satisfies* $\mathrm{NEGap}(\mu^k, \nu^k) = \Omega(\sqrt{XA + YB}k^{-1/2})$*, when* $k \geqslant \max(XA, YB)$.

Proof Sketch. The idea of the proof is to leverage the fact that if an uncoupled algorithm can learn the NE of IIEFGs with a last-iterate convergence guarantee of $\widetilde{\Theta}(k^{-\alpha})$ ($\alpha \in [0, 1]$) in the bandit feedback setting, then it can be used to learn IIEFGs where the opponent is an adversary with a regret of order $\widetilde{\Theta}(k^{1-\alpha})$. Therefore, considering that the hardness of minimizing regret of IIEFGs with an adversarial opponent is equivalent to minimizing regret on a bandit problem with AX arms (Bai et al., 2022; Fiegel et al., 2023), the proof of Theorem 5.3 can be completed by contradiction. ☐

Remark 5.4. *Compared with the lower bound of the convergence rate above, the upper bound in Theorem 5.1 is loose by a factor of* $\widetilde{\mathcal{O}}((X + Y)k^{3/8})$ *(for large enough* X, Y, A *and* B*). We believe one of the promising approaches to improve the upper bound of the convergence rate might be to consider using the optimistic OMD/FTRL, which utilizes accelerated techniques from the optimization perspective and is typically used to achieve the* $\widetilde{\mathcal{O}}(1/k)$ *convergence rate for learning IIEFGs with last-iterate convergence in the full-information setting. One of the main difficulties of using optimistic OMD/FTRL in conjunction with the regularization technique to achieve a faster last-iterate convergence rate of learning IIEFGs in the bandit feedback setting is that the loss estimator constructed in the bandit feedback setting (either unbiased or optimistically biased) to serve as a surrogate of the true loss would have undesirably large variance, making the stability of optimistic OMD/FTRL hard to be controlled even in the special case of learning matrix games. We leave the possible improvement of our convergence upper bound as our future study.*

6 Conclustion

In this work, we make the first step to establishing the algorithm that learns an approximate NE of IIEFGs in the bandit feedback setting with finite-time last-iterate convergence. Our algorithm is fully uncoupled between the two players involved in the games and does not require any coordination, communication, or shared randomness between these players. We prove that our algorithm achieves the last-iterate convergence of order $\widetilde{\mathcal{O}}((X + Y)[(XA + YB)^{1/2} + (X + Y)^{1/4}H]k^{-1/8})$ with high probability and of order $\widetilde{\mathcal{O}}((X+Y)[(X^2A+Y^2B)^{1/2}+(X+Y)^{1/4}H]k^{-1/6})$ in expectation (for large enough k). Also, we provide the lower bound of order $\Omega(\sqrt{XA + YB}k^{-1/2})$ for learning IIEFGs with last-iterate convergence guarantee in the bandit feedback setting. An interesting problem might be closing the gap between the established convergence upper and lower bound, which still remains open in the special case of learning matrix games with the last-iterate convergence guarantee in the bandit feedback setting. We will leave the investigation of this for our future research endeavors.

REFERENCES

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. Last-iterate convergence with full and noisy feedback in two-player zero-sum games. In *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7999–8028. PMLR, 2023.

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, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 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.

Martino Bernasconi, Alberto Marchesi, and Francesco Trovò. Learning extensive-form perfect equilibria in two-player zero-sum sequential games. In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2152–2160. PMLR, 2024.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting CFR+ and alternating updates. *J. Artif. Intell. Res.*, 64:429–443, 2019.

Yang Cai and Weiqiang Zheng. Doubly optimal no-regret learning in monotone 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. 3507–3524. PMLR, 2023.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 27952–27964, 2021.

Shicong Cen, Yuejie Chi, Simon Shaolei Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum markov games. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Ziyi Chen, Shaocong Ma, and Yi Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum markov game. In *International Conference on Learning Representation*, 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.

Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Improved rates for derivative free gradient play in strongly monotone games. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 3403–3408. IEEE, 2022.

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.

Yi Feng, Hu Fu, Qun Hu, Ping Li, Ioannis Panageas, Bo Peng, and Xiao Wang. On the last-iterate convergence in time-varying zero-sum games: Extra gradient succeeds where optimism fails. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

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.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Eduard Gorbunov, Adrien B. Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 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.

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, 2010.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6936–6946, 2019.

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

Yuanhanqing Huang and Jianghai Hu. Zeroth-order learning in continuous games via residual pseudogradient estimates. *arXiv preprint arXiv:2301.02279*, 2023.

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

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.

Michael Jordan, Tianyi Lin, and Zhengyuan Zhou. Adaptive, doubly optimal no-regret learning in strongly monotone and exp-concave games with gradient feedback. *Operations Research*, 2024.

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, 2015.

Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Solving large sequential games with the excessive gap technique. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 872–882, 2018.

Harold W Kuhn. Extensive games. *Proceedings of the National Academy of Sciences*, 36(10): 570–576, 1950.

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.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

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.

Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael I. Jordan. Finite-time last-iterate convergence for multi-agent learning in 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. 6161–6171. PMLR, 2020.

Tianyi Lin, Zhengyuan Zhou, Wenjia Ba, and Jiawei Zhang. Doubly optimal no-regret online learning in strongly monotone games with bandit feedback. *arXiv preprint arXiv:2112.02856*, 2021.

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.

Mingyang Liu, Asuman E. Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

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.

Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM J. Optim.*, 30(4):3230–3251, 2020.

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

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3168–3176, 2015.

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.

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

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.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4259–4299. PMLR, 2021a.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, pp. 4259–4299. PMLR, 2021b.

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.

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.

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 More Discussions on Virtual Transition Probabilities

A.1 Illustration on the Failure of Using Uniform Virtual Transition

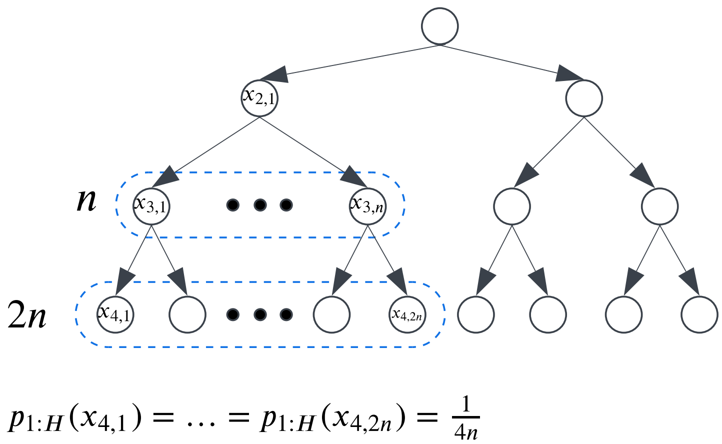

Figure 1: An illustrative example where using uniform virtual transition p fails to guarantee $\min_{x_h \in \mathcal{X}_h, h\in[H]} p_{1:h}(x_h) \geq 1/X$.

On the IIEFG instance shown in Figure 1, there is only one action a and $H = 4$. Each infoset x in the game tree of this instance satisfies $|C(x,a)| = 2$ except for infoset $x_{2,1}$, which is such that $|C(x_{2,1},a)| = n$ with some $n \geq 2$. Now suppose the uniform distribution p is used as a virtual transition over infoset-action spaces. Then for all the descendants $\{x_{4,i}\}_{i=1}^{2n}$ on step $h=4$ of infoset $x_{2,1}$, one can see that $p_{1:H}(x_{H,i}) = \frac{1}{2}\cdot\frac{1}{n}\cdot\frac{1}{2} = \frac{1}{4n}$, while there are only $X = 9+3n$ infosets in total. Thus, it will happen that $p_{1:H}(x_{H,i}) < \frac{1}{X}$ when $n > 9$.

Actually, one can easily construct an IIEFG instance such that $\min_{x_H\in\mathcal{X}_H} p_{1:H}(x_H) \leq \mathcal{O}(\frac{1}{n^m})$ and $X = \mathcal{O}(mn+c)$ with c as a parameter that depends on m but not n for uniform virtual transition p. Therefore, when using uniform distribution p as a virtual transition, $\max_{x_H\in\mathcal{X}_H} 1/p_{1:H}(x_H)$ might be prohibitively large and lead to a convergence rate with much worse dependence on X than the virtual transition constructed in our Algorithm 2.

A.2 Balanced Effects of the Proposed Virtual Transition Probability

Lemma A.1. *For any $h\in[H]$ and $x_h\in\mathcal{X}_h$, the constructed virtual transition p^x guarantees that $1/p^x_{1:h}(x_h) \leq X$.*

Proof. Clearly, $p^x_{1:h}(\cdot)$ is minimzed at $h = H$ for some $x_H \in \mathcal{X}_H$ by the definition of virtual transition. By the construction of $p^x_{1:h}(\cdot)$ in Algorithm 2, one can deduce that $\forall x_H\in\mathcal{X}_H$, it holds that (understanding $\{(x_h,a_h)\}_{h\in[H-1]}$ as the unique trajectory leading to x_H below)

$$\begin{aligned}
p^x_{1:H}(x_H) &= q[x_H]\\
&= q\left[x_{H-1}\right]\cdot\frac{c\left[x_H\right]}{\sum_{x'_H\in C(x_{H-1},a_{H-1})} c\left[x'_H\right]}\\
&= q\left[x_{H-2}\right]\cdot\frac{c\left[x_{H-1}\right]}{\sum_{x'_{H-1}\in C(x_{H-2},a_{H-2})} c\left[x'_{H-1}\right]}\cdot\frac{c\left[x_H\right]}{\sum_{x'_H\in C(x_{H-1},a_{H-1})} c\left[x'_H\right]}\\
&= q\left[x_{H-2}\right]\cdot\frac{c\left[x_{H-1}\right]}{\sum_{x'_{H-1}\in C(x_{H-2},a_{H-2})} c\left[x'_{H-1}\right]}\cdot\frac{c\left[x_H\right]}{d\left[x_{H-1},a_{H-1}\right]}\\
&\overset{(i)}{\geq} q\left[x_{H-2}\right]\cdot\frac{c\left[x_{H-1}\right]}{\sum_{x'_{H-1}\in C(x_{H-2},a_{H-2})} c\left[x'_{H-1}\right]}\cdot\frac{c\left[x_H\right]}{c[x_{H-1}]}
\end{aligned}$$

$$
\begin{aligned}
&= q\left[x_{H-2}\right] \cdot \frac{c\left[x_H\right]}{\sum_{x'_{H-1} \in C(x_{H-2}, a_{H-2})} c\left[x'_{H-1}\right]} \\
&\geq \ldots \\
&\geq \frac{c\left[x_H\right]}{\sum_{x_1 \in \mathcal{X}_1} c\left[x_1\right]} \\
&\geq \frac{c\left[x_H\right]}{X_H} \\
&\geq \frac{c\left[x_H\right]}{X} \\
&= \frac{1}{X},
\end{aligned}
$$

where $c[\cdot]$, $q[\cdot]$, and $d[\cdot,\cdot]$ are defined in our Algorithm 2; and (i) is due to $c[x_{H-1}] = \max_{a \in \mathcal{A}} d[x_{H-1}, a] \geq d[x_{H-1}, a_{H-1}]$. □

The property shown in this lemma of our constructed virtual transition p^x serves as a key ingredient in the analysis (say, when bounding our **Term 4** and when establishing the final convergence upper bound of the NE gap in the proof of Theorem 5.1) as we shall see.

B Proof of High-probability Last-iterate Convergence Rate

Lemma B.1 (One-step analysis of OMD with virtual transition weighted negentropy regularized loss)**.** *Let*

$$
\begin{aligned}
\mu' = \operatorname{argmin}_{\tilde{\mu} \in \Omega} \sum_{h, x_h, a_h} & \tilde{\mu}_{1:h}\left(x_h, a_h\right)\left(\ell\left(x_h, a_h\right) + \varepsilon\left(x_h, a_h\right) p^x_{1:h}\left(x_h\right) \log\left(p^x_{1:h}\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right)\right) \\
& + \frac{1}{\eta} D_\psi(\tilde{\mu}, \mu),
\end{aligned}
$$

for some convex set $\Omega \subseteq \Pi_{\max}$, $\ell \in \mathbb{R}^{XA}_{\geqslant 0}$, and $\varepsilon \in \left[0, \frac{1}{\eta}\right]^{XA}$. Then $\forall u \in \Omega$.

$$
\begin{aligned}
&\langle \mu' - \mu, \ell + \varepsilon p \log p\mu \rangle \\
\leqslant & \sum_{h, (x_h, a_h)} \left[\frac{\eta}{p^x_{1:h}(x_h)} \mu_{1:h}(x_h, a_h) \ell^2(x_h, a_h) + \eta \varepsilon^2(x_h, a_h) \log^2(p^x_{1:h} \mu_{1:h}(x_h, a_h))\right],
\end{aligned}
$$

where $(\varepsilon p \log p\mu)[(x_h, a_h)] := \varepsilon p^x_{1:h}(x_h) \log\left(p^x_{1:h}(x_h) \mu_{1:h}(x_h, a_h)\right)$.

Proof. The common one-step analysis of OMD shows that

$$
\langle \mu' - \mu, \ell + \varepsilon p \log p\mu \rangle \leqslant \frac{1}{\eta}\left(D_\psi(u, \mu) - D_\psi(u, \mu') - D_\psi(\mu', \mu)\right).
$$

Then, to upper bound $\langle \mu - \mu', l + \varepsilon p \log p\mu \rangle - \frac{1}{\eta} D_\psi\left(\mu', \mu\right)$, notice that

$$
\begin{aligned}
&\langle \mu - \mu', \ell + \varepsilon p \log p\mu \rangle - \frac{1}{\eta} D_\psi\left(\mu', \mu\right) \\
\leqslant & \sup_{v \in \mathbb{R}^{XA}_{\geqslant 0}} \left(\langle \mu - v, \ell + \varepsilon p \log p\mu \rangle - \frac{1}{\eta} D_\psi(v, \mu)\right) \\
= & \langle \mu, \ell + \varepsilon p \log p\mu \rangle - \inf_{v \in \mathbb{R}^{XA}_{\geqslant 0}} \left(\langle v, \ell + \varepsilon p \log p\mu \rangle + \frac{1}{\eta} D_\psi(v, \mu)\right).
\end{aligned}
$$

Further, the first-order optimality condition $\ell + \varepsilon p \log p\mu + \frac{1}{\eta}(\nabla\psi(v) - \nabla\psi(\mu)) = 0$ implies that

$$
\log \frac{v_{1:h}\left(x_h, a_h\right)}{\mu_{1:h}\left(x_h, a_h\right)} = -\frac{\eta}{p^x_{1:h}\left(x_h\right)}\left[\ell\left(x_h, a_h\right) + \varepsilon\left(x_h, a_h\right) p^x_{1:h}\left(x_h\right) \log\left(p^x_{1:h}\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right)\right].
$$

Hence, one can see that

$$
\begin{aligned}
& v_{1:h}\left(x_h, a_h\right) \\
= & \mu_{1:h}\left(x_h, a_h\right) \exp \left(-\frac{\eta}{p_{1:h}^x\left(x_h\right)}\left[\ell\left(x_h, a_h\right)+\varepsilon\left(x_h, a_h\right) p_{1:h}^x\left(x_h\right) \log \left(p_{1:h}^x\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right)\right]\right) .
\end{aligned} \tag{10}
$$

Therefore, we have

$$
\begin{aligned}
& \langle\mu-\mu^{\prime}, \ell+\varepsilon p \log (p \mu)\rangle-\frac{1}{\eta} D_\psi\left(\mu^{\prime}, \mu\right) \\
= & \sum_{h,\left(x_h, a_h\right)}\Big[\left(\mu_{1:h}\left(x_h, a_h\right)-v_{1:h}\left(x_h, a_h\right)\right)\left(\ell\left(x_h, a_h\right)+\varepsilon\left(x_h, a_h\right) p_{1:h}^x\left(x_h\right) \log \left(p_{1:h}^x\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right)\right) \\
& -\frac{1}{\eta}\left(p_{1:h}^x(x_h) v_{1:h}(x_h, a_h) \log \frac{v_{1:h}(x_h, a_h)}{\mu_{1:h}(x_h)}-p_{1:h}^x(x_h)(v_{1:h}(x_h, a_h)-\mu_{1:h}(x_h, a_h))\right)\Big] \\
= & \sum_{h,\left(x_h, a_h\right)}\Big[\left(\mu_{1:h}\left(x_h, a_h\right)\right)\left(\ell\left(x_h, a_h\right)+\varepsilon\left(x_h, a_h\right) p_{1:h}^x\left(x_h\right) \log \left(p_{1:h}^x\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right)\right) \\
& +\frac{p_{1:h}^x(x_h)}{\eta}\left(\exp \left(-\frac{\eta}{p_{1:h}^x(x_h)}[\ell(x_h, a_h)+\varepsilon(x_h, a_h) p_{1:h}^x(x_h) \log (p_{1:h}^x(x_h) \mu_{1:h}(x_h, a_h))]\right)-1\right) \mu_{1:h}(x_h, a_h)\Big] \\
= & \sum_{h,\left(x_h, a_h\right)} \frac{p_{1:h}^x(x_h)}{\eta} \mu_{1:h}(x_h, a_h)\Big[\frac{\eta}{p_{1:h}^x(x_h)}\left(\ell\left(x_h, a_h\right)+\varepsilon\left(x_h, a_h\right) p_{1:h}^x\left(x_h\right) \log \left(p_{1:h}^x\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right)\right) \\
& +\left(\exp \left(-\frac{\eta}{p_{1:h}^x(x_h)}[\ell(x_h, a_h)+\varepsilon(x_h, a_h) p_{1:h}^x(x_h) \log (p_{1:h}^x(x_h) \mu_{1:h}(x_h, a_h))]\right)-1\right)\Big] \\
\leqslant & \sum_{h,\left(x_h, a_h\right)} \frac{\eta}{p_{1:h}^x(x_h)} \mu_{1:h}(x_h, a_h) \ell^2(x_h, a_h) \\
& +\sum_{h,\left(x_h, a_h\right)} \frac{p_{1:h}^x(x_h)}{\eta}\Big[\mu_{1:h}\left(x_h, a_h\right)\left(\ell\left(x_h, a_h\right)+\varepsilon\left(x_h, a_h\right) p_{1:h}^x\left(x_h\right) \log \left(p_{1:h}^x\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right)\right) \\
& +\mu_{1:h}(x_h, a_h) \exp \left(-\frac{\eta}{p_{1:h}^x(x_h)}[\ell(x_h, a_h)+\varepsilon(x_h, a_h) p_{1:h}^x(x_h) \log (p_{1:h}^x(x_h) \mu_{1:h}(x_h, a_h))]\right) \\
& -\mu_{1:h}(x_h, a_h) \exp \left(-\frac{\eta}{p_{1:h}^x(x_h)} \ell(x_h, a_h)\right)\Big] \\
= & \sum_{h,\left(x_h, a_h\right)} \frac{\eta}{p_{1:h}^x(x_h)} \mu_{1:h}(x_h, a_h) \ell^2(x_h, a_h) \\
& +\sum_{h,\left(x_h, a_h\right)} \frac{1}{\eta}\Big[\mu_{1:h}\left(x_h, a_h\right) \eta \varepsilon\left(x_h, a_h\right) p_{1:h}^x\left(x_h\right) \log \left(p_{1:h}^x\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right) \\
& +\exp \left(-\frac{\eta}{p_{1:h}^x} \ell(x_h, a_h)\right)\left((p_{1:h}^x \mu_{1:h}(x_h, a_h))^{1-\eta \varepsilon(x_h, a_h)}-p_{1:h}^x \mu_{1:h}(x_h, a_h)\right)\Big] \\
\leqslant & \sum_{h,\left(x_h, a_h\right)} \frac{\eta}{p_{1:h}^x(x_h)} \mu_{1:h}(x_h, a_h) \ell^2(x_h, a_h) \\
& +\sum_{h,\left(x_h, a_h\right)} \frac{1}{\eta}\Big[\mu_{1:h}\left(x_h, a_h\right) \eta \varepsilon\left(x_h, a_h\right) p_{1:h}^x\left(x_h\right) \log \left(p_{1:h}^x\left(x_h\right) \mu_{1:h}\left(x_h, a_h\right)\right) \\
& -\eta \varepsilon(x_h, a_h)(p_{1:h}^x(x_h) \mu_{1:h}(x_h, a_h))^{1-\eta \varepsilon(x_h, a_h)} \ln (p_{1:h}^x(x_h) \mu_{1:h}(x_h, a_h))\Big] \\
\leqslant & \sum_{h,\left(x_h, a_h\right)}\left[\frac{\eta}{p_{1:h}^x(x_h)} \mu_{1:h}(x_h, a_h) \ell^2(x_h, a_h)+\eta \varepsilon^2(x_h, a_h) \log ^2(p_{1:h}^x(x_h) \mu_{1:h}(x_h, a_h))\right],
\end{aligned}
$$

where in the second equality we substitute $v_{1:h}(x_h, a_h)$ with Eq. (10), in first inequality comes from the fact that $\frac{\eta}{p_{1:h}^x(x_h)}\ell\left(x_h, a_h\right) \leqslant \left(\eta\ell\left(x_h, a_h\right)/p_{1:h}^x(x_h)\right)^2 - \exp(\eta\ell\left(x_h, a_h\right)/p_{1:h}^x(x_h))$ and the forth equality follows from

$$\begin{aligned}&\exp\left(-\frac{\eta}{p^x_{1:h}(x_h)}\left[\ell(x_h,a_h)+\varepsilon(x_h,a_h)p^x_{1:h}(x_h)\log(p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))\right]\right)\\=&\exp\left(-\frac{\eta}{p^x_{1:h}(x_h)}\ell(x_h,a_h)\right)\left((p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h)\right)^{1-\eta\varepsilon(x_h,a_h)}\ .\end{aligned}$$

The last two inequalities can be derived by following calculations:

$$\begin{aligned}&\exp\left(-\frac{\eta}{p^x_{1:h}(x_h)}\ell(x_h,a_h)\right)\left((p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))^{1-\eta\varepsilon(x_h,a_h)}-p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h)\right)\\\leqslant&(p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))^{1-\eta\varepsilon(x_h,a_h)}-p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h)\\\leqslant&\eta\varepsilon(x_h,a_h)(p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))^{1-\eta\varepsilon(x_h,a_h)}\ln(p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))\ ,\end{aligned}$$

and

$$\begin{aligned}&\mu_{1:h}\left(x_h,a_h\right)\eta\varepsilon\left(x_h,a_h\right)p^x_{1:h}\left(x_h\right)\log\left(p^x_{1:h}\left(x_h\right)\mu_{1:h}\left(x_h,a_h\right)\right)\\&-\eta\varepsilon(x_h,a_h)(p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))^{1-\eta\varepsilon(x_h,a_h)}\ln(p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))\\=&-\eta\varepsilon(x_h,a_h)\log(p^x_{1:h}\left(x_h\right)\mu_{1:h}(x_h,a_h))((p^x_{1:h}\left(x_h\right)\mu_{1:h}(x_h,a_h))^{1-\eta\varepsilon(x_h,a_h)}-p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h))\\\leqslant&\eta^2\varepsilon^2(x_h,a_h)(\log^2(p^x_{1:h}\left(x_h\right)\mu_{1:h}(x_h,a_h))^{1-\eta\varepsilon(x_h,a_h)}\\\leqslant&\eta^2\varepsilon^2(x_h,a_h)(\log^2(p^x_{1:h}\left(x_h\right)\mu_{1:h}(x_h,a_h))\ .\end{aligned}$$

□

Lemma B.2. $\forall k\geqslant 1$*, we have*

$$\mathrm{NEGap}(\xi^{k,*})=\mathcal{O}\left(\varepsilon_k H(\ln(XA)+\ln(YB))+\frac{XAH}{k}+\frac{YBH}{k}\right)\ . \tag{11}$$

Proof. $\forall\left(\mu',\nu'\right)\in\Pi_{\max}\times\Pi_{\min}$, we have

$$\begin{aligned}&f\left(\mu^{k,\star},\nu'\right)-f\left(\mu',\nu^{k,\star}\right)\\=&f\left(\mu^{k,\star},\nu'\right)-f\left(\mu^{k,\star},\nu\right)+\left(\mu^{k,\star},\nu\right)-f\left(\mu,\nu^{k,\star}\right)+f\left(\mu,\nu^{k,\star}\right)-f\left(\mu',\nu^{k,\star}\right)\ .\end{aligned}$$

First notice that $\forall(\mu,v)\in\Pi^k_{\max}\times\Pi^k_{\min}$,

$$\begin{aligned}&f\left(\mu^{k,\star},\nu\right)-f\left(\mu,\nu^{k,\star}\right)\\=&f\left(\mu^{k,\star},\nu\right)-f_k\left(\mu^{k,\star},\nu\right)+f_k\left(\mu^{k,\star},\nu\right)-f_k\left(\mu,\nu^{k,\star}\right)+f_k\left(\mu,\nu^{k,\star}\right)-f\left(\mu,\nu^{k,\star}\right)\\=&-\left(\varepsilon_k\psi\left(\mu^{k,\star}\right)-\varepsilon_k\psi(\nu)\right)+\left(\varepsilon_k\psi(\mu)-\varepsilon_k\psi\left(\nu^{k,\star}\right)\right)\\\leqslant&-\varepsilon_k\psi\left(\mu^{k,\star}\right)-\varepsilon_k\psi\left(\nu^{k,\star}\right)\\\leqslant&\varepsilon_k H(\ln(XA)+\ln(YB))\ .\end{aligned}$$

To bound $f\left(\mu^{k,\star},\nu'\right)-f\left(\mu^{k,\star},\nu\right)$, we have

$$\begin{aligned}&f\left(\mu^{k,\star},\nu'\right)-f\left(\mu^{k,\star},\nu\right)\\\leqslant&\left\langle\nabla_\nu f\left(\mu^{k,\star},\nu\right),\nu'-\nu\right\rangle\\\leqslant&\|\nabla_\nu f\left(\mu^{k,\star},\nu\right)\|_1\|\nu'-\nu\|_\infty\\\leqslant&YB\left(1-\left(1-\frac{B-1}{Bk}\right)^H\right)\\\leqslant&YB\left(1-\left(1-\frac{B}{Bk}\right)^H\right)\\\leqslant&YB\left(1-\left(1-\frac{1}{k}\right)^H\right)\end{aligned}$$

$$=\mathcal{O}\left(\frac{YBH}{k}\right).$$

Similarly, we have

$$f\left(\mu,\nu^{k,\star}\right)-f\left(\mu',\nu^{k,\star}\right)\leqslant\mathcal{O}\left(\frac{XAH}{k}\right).$$

Putting all the above together completes the proof. □

B.1 Convergence Rate of the Contraction of the Bregman Divergence

Lemma B.3 (Contraction on Bregman divergence)**.**

$$\begin{aligned}
&D_\psi\left(\xi^{k+1,\star},\xi^{k+1}\right)\leqslant\underbrace{\sum_{i=1}^{k}w_k^i\eta_i\rho_i}_{\textbf{Term 1}}+\underbrace{\sum_{i=1}^{k}w_k^i\eta_i\sigma_i}_{\textbf{Term 2}}\\
&+\underbrace{XA\left(\log X+H\log\left(Ak\right)\right)^2\sum_{i=1}^{k}w_k^i\left(\eta_i\varepsilon_i\right)^2+YB\left(\log Y+H\log\left(Bk\right)\right)^2\sum_{i=1}^{k}w_k^i\left(\eta_i\varepsilon_i\right)^2}_{\textbf{Term 3}}\\
&+\underbrace{\sum_{i=1}^{k}w_k^i\eta_i^2\left(X\underline{\tau}_i+Y\bar{\tau}_i\right)}_{\textbf{Term 4}}+\underbrace{\sum_{i=1}^{k}w_k^i\eta_i^2\left(X^2A+Y^2B\right)}_{\textbf{Term 5}}+\underbrace{\sum_{i=1}^{k}w_k^i\omega_i}_{\textbf{Term 6}}.
\end{aligned}$$

Proof. Recall we denote $[p^x\mu](x_h,a_h)\coloneqq p^x_{1:h}(x_h)\mu_{1:h}(x_h,a_h)$.

$$f_k(\mu^k,\nu^k)-f_k(\mu^{k,\star},\nu^k)=\left(\mu^k-\mu^{k,\star}\right)^\top\boldsymbol{G}\nu^k+\varepsilon_k\left(\psi\left(\mu^k\right)-\psi\left(\mu^{k,\star}\right)\right).$$

For the first term in the above display, we have

$$\begin{aligned}
&\left(\mu^k-\mu^{k,\star}\right)^\top\boldsymbol{G}\nu^k\\
=&\left(\mu^k-\mu^{k,\star}\right)^\top\left(\boldsymbol{G}\nu^k+g^k-g^k\right)\\
=&\left(\mu^k-\mu^{k,\star}\right)^\top g^k+\left(\mu^k\right)^\top\left(\boldsymbol{G}\nu^k-g^k\right)-\left(\mu^{k,\star}\right)^\top\left(\boldsymbol{G}\nu^k-g^k\right)\\
=&\left(\mu^k-\mu^{k,\star}\right)^\top g^k+\sum_{h,x_h,a_h}\mu^k_{1:k}\left(x_h,a_h\right)\left[\left(\boldsymbol{G}\nu^k\right)\left[\left(x_h,a_h\right)\right]-g^k\left[\left(x_h,a_h\right)\right]\right]\\
&-\sum_{h,x_h,a_h}\mu^{k,\star}_{1:h}\left(x_n,a_n\right)\left[\left(\boldsymbol{G}\nu^k\right)\left[\left(x_h,a_h\right)\right]-g^k\left[\left(x_h,a_h\right)\right]\right]\\
=&\left(\mu^k-\mu^{k,\star}\right)^\top g^k\\
&+\sum_{h,x_h,a_h}\mu^k_{1:h}\left(x_h,a_h\right)\left[\left(\boldsymbol{G}\nu^k\right)\left[\left(x_h,a_h\right)\right]-\left(\frac{\mathbb{I}^k_h\left\{x_h,a_h\right\}}{\mu^k_{1:h}\left(x_h,a_h\right)+\gamma_k}\left(1-r^k_h\right)+\varepsilon_k p^x_{1:h}\left(x_h\right)\log\left[p^x\mu^k\right]\left(x_n,a_n\right)\right)\right]\\
&-\sum_{h,x_h,a_h}\mu^{k,\star}_{1:h}\left(x_h,a_h\right)\left[\left(\boldsymbol{G}\nu^k\right)\left[\left(x_h,a_h\right)\right]-\left(\frac{\mathbb{I}^k_h\left\{x_h,a_h\right\}}{\mu^k_{1:h}\left(x_h,a_h\right)+\gamma_k}\left(1-r^k_h\right)+\varepsilon_k p^x_{1:h}\left(x_h\right)\log\left[p^x\mu^k\right]\left(x_n,a_n\right)\right)\right].
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\psi\left(\mu^k\right)-\psi\left(\mu^{k,\star}\right)\\
=&\sum_{h,x_h,a_h}\left[p^x\mu^k\right]\left(x_h,a_h\right)\log\left[p^x\mu^k\right]\left(x_h,a_h\right)\\
&-\sum_{h,x_h,a_h}\left[p^x\mu^{k,\star}\right]\left(x_h,a_h\right)\log\left[p^x\mu^{k,\star}\right]\left(x_h,a_h\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h,x_h,a_h} \left(\left[p^x \mu^k\right](x_h,a_h) - \left[p^x \mu^{k,\star}\right](x_h,a_h)\right) \log\left[p^x \mu^k\right](x_h,a_h) \\
&\quad - \sum_{h,x_h,a_h} \left[p^x \mu^{k,\star}\right](x_h,a_h) \left(\log\left[p^x \mu^{k,\star}\right](x_h,a_h) - \log\left[p^x \mu^k\right](x_h,a_h)\right) \\
&= \sum_{h,x_h,a_h} \left(\left[p^x \mu^k\right](x_h,a_h) - \left[p^x \mu^{k,\star}\right](x_h,a_h)\right) \log\left[p^x \mu^k\right](x_h,a_h) - D_\psi(\mu^{k,\star},\mu^k)\,.
\end{aligned}
$$

We then arrive at

$$
\begin{aligned}
& f_k(\mu_k, v_k) - f_k(\mu_k^\star, v_k) \\
&= \left(\mu^k - \mu^{k,\star}\right)^\top g^k + \underbrace{\sum_{h,x_h,a_h} \mu^k_{1:k}(x_h,a_h)\left[\left(\boldsymbol{G}\nu^k\right)[(x_h,a_h)] - \left(\frac{\mathbb{I}_h^k\{x_h,a_h\}}{\mu^k_{1:h}(x_h,a_h)+\gamma_k}\left(1-r_h^k\right)\right)\right]}_{=:\underline{\rho_k}} \\
&\quad \underbrace{- \sum_{h,x_h,a_h} \mu^{k,\star}_{1:k}(x_h,a_h)\left[\left(\boldsymbol{G}\nu^k\right)[(x_h,a_h)] - \left(\frac{\mathbb{I}_h^k\{x_h,a_h\}}{\mu^k_{1:h}(x_h,a_h)+\gamma_k}\left(1-r_h^k\right)\right)\right]}_{=:\underline{\sigma_k}} - \varepsilon_k D_\psi(\mu^{k,\star},\mu^k) \\
&\leqslant \frac{1}{\eta_k}\left(D_\psi(\mu^{k,\star},\mu^k) - D_\psi(\mu^{k,\star},\mu^{k+1})\right) - \varepsilon_k D_\psi(\mu^{k,\star},\mu^k) + \underline{\rho_k} + \underline{\sigma_k} \\
&\quad + \sum_{h,x_h,a_h} \eta_k \left(\frac{1}{p^x_{1:h}(x_h)} \mu^k_{1:h}(x_h,a_h) \hat{\ell}^k_h(x_h,a_h)^2 + \varepsilon_k^2(x_h,a_h)\log^2(p^x_{1:h}(x_h)\mu^k_{1:h}(x_h,a_h)\right) \\
&\leqslant \frac{(1-\eta_k\varepsilon_k) D_\psi\left(\mu^{k,\star},\mu^k\right) - D_\psi\left(\mu^{k,\star},\mu^{k+1}\right)}{\eta_k} + \underline{\rho_k} + \underline{\sigma_k} \\
&\quad + \sum_{h,x_h,a_h} \eta_k \left(\frac{1}{p^x_{1:h}(x_h)} \frac{\mathbb{I}_h^k\{x_h,a_h\}}{\mu^k_{1:h}(x_h,a_h)+\gamma_k} + \varepsilon_k^2 \log^2(\underbrace{p^x_{1:h}(x_h)\mu^k_{1:h}(x_h,a_h)}_{m:=\min_{h,(x_h,a_h)} p^x_{1:h}(x_h)\mu^k_{1:h}(x_h,a_h)})\right) \\
&\leqslant \frac{(1-\eta_k\varepsilon_k) D_\psi\left(\mu^{k,\star},\mu^k\right) - D_\psi\left(\mu^{k,\star},\mu^{k+1}\right)}{\eta_k} + \underline{\rho_k} + \underline{\sigma_k} \\
&\quad + \eta_k \underbrace{\sum_{h,x_h,a_h} \frac{1}{p^x_{1:h}(x_h)} \left(\frac{\mathbb{I}_h^k\{x_h,a_h\}}{\mu^k_{1:h}(x_h,a_h)+\gamma_k} - 1\right)}_{=:X\underline{\tau_k}} + \eta_k X^2 A + \eta_k X A \log^2 m \\
&\leqslant \frac{(1-\eta_k\varepsilon_k) D_\psi\left(\mu^{k,\star},\mu^k\right) - D_\psi\left(\mu^{k,\star},\mu^{k+1}\right)}{\eta_k} + \underline{\rho_k} + \underline{\sigma_k} + \eta_k X \underline{\tau_k} + \eta_k X^2 A + \eta_k X A \log^2 m\,.
\end{aligned}
$$

Rearranging shows that

$$
\begin{aligned}
& D_\psi\left(\mu^{k+1,\star},\mu^{k+1}\right) \\
&\leqslant (1-\eta_k\varepsilon_k) D_\psi\left(\mu^{k,\star},\mu^k\right) + \eta_k\left(f_k\left(\mu^{k,\star},\nu_k\right) - f_k\left(\mu_k,\nu_k\right)\right) \\
&\quad + \eta_k^2 X A \log^2 m + \eta_k^2 X A \underline{\tau_k} + \eta_k^2 X^2 A + \eta_k \underline{\rho_k} + \eta_k \underline{\sigma_k} + \underbrace{D_\psi\left(\mu^{k+1,\star},\mu^{k+1}\right) - D_\psi\left(\mu^{k,\star},\mu^{k+1}\right)}_{=:\underline{\omega_k}}\,.
\end{aligned}
$$

Analogously, for the min-player, we have

$$
\begin{aligned}
& D_\psi\left(\nu^{k+1,\star},\nu^{k+1}\right) \\
&\leqslant (1-\eta_k\varepsilon_k) D_\psi\left(\nu^{k,\star},\nu^k\right) + \eta_k\left(f_k\left(\mu^k,\nu^k\right) - f_k\left(\mu_k,\nu^{k,\star}\right)\right) \\
&\quad + \eta_k^2 Y B \log^2 m + \eta_k^2 Y \bar{\tau}_k + \eta_k^2 Y^2 B + \eta_k \bar{\rho}_k + \eta_k \bar{\sigma}_k + \bar{w}_k\,,
\end{aligned}
$$

where

$$
\bar{\tau}_k := \frac{1}{Y} \sum_{h,y_h,b_h} \frac{1}{p^y_{1:h}(y_h)} \left(\frac{\mathbb{I}_h^k\{y_h,b_h\}}{\nu^k_{1:h}(y_h,b_h)+\gamma_k} - 1\right)
$$

$$\bar{\rho}_k \coloneqq \sum_{h,y_h,b_h} \nu_{1:h}^k(y_h,b_h)\left[\left(1-\left(\boldsymbol{G}^\top\mu^k\right)[(y_h,b_h)]\right)-\frac{\mathbb{I}_h^k\{y_h,b_h\}\,r_h^k}{\nu_{1:h}^k(y_h,b_h)+\gamma_{k\,k}}\right]$$
$$\bar{\sigma}_k \coloneqq \sum_{h,y_h,b_h} \nu_{1:h}^{k,\star}(y_h,b_h)\left[\frac{\mathbb{I}_h^k\{y_h,b_h\}\,r_h^k}{\nu_{1:h}^k(y_h,b_h)+\gamma_{k\,k}}-\left(1-\left(\boldsymbol{G}^\top\mu^k\right)[(y_h,b_h)]\right)\right]$$
$$\bar{\omega}_k \coloneqq D_\psi\left(\nu^{k+1,\star},\nu^{k+1}\right)-D_\psi\left(\nu^{k,\star},\nu^{k+1}\right)\,.$$

Combining both sides and noticing that $f_k\left(\mu^{k,\star},\nu^k\right)-f_k\left(\mu^k,\nu^{k,\star}\right)\leqslant 0$, we have

$$\begin{aligned}
&D_\psi\left(\xi^{k+1,x},\xi^{k+1}\right)\\
&\leqslant(1-\eta_k\varepsilon_k)\,D_\psi\left(\xi^{k,\star},\xi^k\right)+\eta_k^2\left(X\underline{\tau}_k+Y\bar{\tau}_k\right)+\eta_k^2\left(X^2A+Y^2B\right)+\eta_k\rho_k+\eta_k\sigma_k+\omega_k\\
&\quad+\eta_k^2XA\varepsilon_k^2\left(\log X+H\log(Ak)\right)^2+\eta_k^2YB\varepsilon_k^2\left(\log Y+H\log(Bk)\right)^2\,.
\end{aligned}$$

Now expanding the recursion in the above display leads to

$$\begin{aligned}
D_\psi\left(\xi^{k+1,\star},\xi^{k+1}\right)\leqslant{}&\underbrace{\sum_{i=1}^k w_k^i\eta_i\rho_i}_{\textbf{Term 1}}+\underbrace{\sum_{i=1}^k w_k^i\eta_i\sigma_i}_{\textbf{Term 2}}\\
&+\underbrace{XA\left(\log X+H\log(Ak)\right)^2\sum_{i=1}^k w_k^i\left(\eta_i\varepsilon_i\right)^2+YB\left(\log Y+H\log(Bk)\right)^2\sum_{i=1}^k w_k^i\left(\eta_i\varepsilon_i\right)^2}_{\textbf{Term 3}}\\
&+\underbrace{\sum_{i=1}^k w_k^i\eta_i^2\left(X\underline{\tau}_i+Y\bar{\tau}_i\right)}_{\textbf{Term 4}}+\underbrace{\sum_{i=1}^k w_k^i\eta_i^2\left(X^2A+Y^2B\right)}_{\textbf{Term 5}}+\underbrace{\sum_{i=1}^k w_k^i\omega_i}_{\textbf{Term 6}}\,,
\end{aligned}$$

where $w_k^i=\prod_{j=i+1}^k(1-\eta_j\varepsilon_j)$. □

B.2 Bounding Contraction Terms

Lemma B.4 (Bounding **Term 1**).

$$\textbf{Term 1}\leqslant(XA+YB)\ln(k)k^{-\alpha_{\gamma_k}+\alpha_\varepsilon}+k^{-\frac{\alpha_k}{2}+\frac{\alpha_\varepsilon}{2}}\log\left(\frac{k^2}{\delta}\right)\,.$$

Proof. Recall

$$\textbf{Term 1}=\sum_{i=1}^k w_k^i\eta_i\rho_i=\sum_{i=1}^k w_k^i\eta_i\underline{\rho}_i+\sum_{i=1}^k w_k^i\eta_i\bar{\rho}_i\,.$$

To bound $\sum_{i=1}^k w_k^i\eta_i\underline{\rho}_i$, note that

$$\begin{aligned}
&\sum_{i=1}^k w_k^i\eta_i\underline{\rho}_i\\
&=\sum_{i=1}^k w_k^i\eta_i\left\langle\mu^i,\ell^{i,x}-\hat{\ell}^{i,x}\right\rangle\\
&=XA\sum_{i=1}^k w_k^i\eta_i\gamma_{k\,i}+H\sqrt{2\sum_{i=1}^k\left(w_k^i\eta_i\right)^2\log\frac{k^2}{\delta}}
\end{aligned}$$

$$
\begin{aligned}
&\leqslant XA \sum_{i=1}^{k} \left[i^{-\alpha_{\gamma_k}-\alpha_\eta} \prod_{j=i+1}^{k} \left(1 - j^{-\alpha_\eta - \alpha_\varepsilon}\right) \right] + \sqrt{ \log\left(\frac{k^2}{\delta}\right) \sum_{i=1}^{k} \left[i^{-2\alpha_\eta} \left(\prod_{j=i+1}^{k} \left(1 - j^{-\alpha_\gamma - \alpha_\varepsilon}\right) \right)^2 \right] } \\
&\leqslant XA \sum_{i=1}^{k} \left[i^{-\alpha_{\gamma}-\alpha_\eta} \prod_{j=i+1}^{k} \left(1 - j^{-\alpha_n - \alpha_\varepsilon}\right) \right] + \sqrt{ \log\left(\frac{k^2}{\delta}\right) \sum_{i=1}^{k} \left[i^{-2\alpha_\eta} \left(\prod_{j=i+1}^{k} \left(1 - j^{-\alpha_n - \alpha_\varepsilon}\right) \right) \right] } \\
&\leqslant XA \ln(k) k^{-\alpha_\gamma + \alpha_\varepsilon} + \sqrt{ \log\left(\frac{k^2}{\delta}\right) \ln(k) k^{-\alpha_\gamma + \alpha_\varepsilon} } \\
&\leqslant XA \ln(k) k^{-\alpha_\gamma + \alpha_\varepsilon} + k^{-\frac{\alpha_\eta}{2} + \frac{\alpha_\varepsilon}{2}} \log\left(\frac{k^2}{\delta}\right) ,
\end{aligned}
$$

where the second equality is given by Lemma B.13 and the third inequality comes from Lemma E.1.

Analogously, we have

$$
\sum_{i=1}^{k} w_k^i \eta_i \bar{\rho}_i \leqslant YB \ln(k) k^{-\alpha_\gamma + \alpha_\varepsilon} + k^{-\frac{\alpha_k}{2} + \frac{\alpha_\varepsilon}{2}} \log\left(\frac{k^2}{\delta}\right) .
$$

Hence

$$
\textbf{Term 1} \leqslant (XA + YB) \ln(k) k^{-\alpha_\gamma + \alpha_\varepsilon} + k^{-\frac{\alpha_k}{2} + \frac{\alpha_\varepsilon}{2}} \log\left(\frac{k^2}{\delta}\right) .
$$

□

Lemma B.5 (Bounding **Term 2**).

$$
\textbf{Term 2} \leqslant k^{-\alpha_\eta + \alpha_{\gamma_k}} \log \frac{k^2}{\delta} .
$$

Proof.

$$
\begin{aligned}
\textbf{Term 2} &= \sum_{i=1}^{k} w_k^i \eta_i \sigma_i \\
&= \sum_{i=1}^{k} w_k^i \eta_i \underline{\sigma}_i + \sum_{i=1}^{k} w_k^i \eta_i \bar{\sigma}_i \\
&\leqslant \max_{1 \leqslant i \leqslant k} \frac{\eta_i w_k^i}{\gamma_{k k}} \log \frac{k^2}{\delta} \quad (\text{with probability } 1 - \frac{k^2}{\delta}) \\
&\leqslant k^{-\alpha_\eta + \alpha_{\gamma_k}} \log \frac{k^2}{\delta} ,
\end{aligned}
$$

where the last inequality is due to Lemma E.2. □

Lemma B.6 (Bounding **Term 3**).

$$
\textbf{Term 3} \leqslant \left(XA \left(\log X + H \log (Ak)^2 \right) + YB \left(\log Y + H \log (Bk) \right)^2 \right) k^{-\alpha_\eta - \alpha_\varepsilon} .
$$

Proof.

$$
\begin{aligned}
&\textbf{Term 3} \\
=& XA \left(\log X + H \log (Ak) \right)^2 \sum_{i=1}^{k} w_k^i \left(\eta_i \varepsilon_i\right)^2 + YB \left(\log Y + H \log (Bk) \right)^2 \sum_{i=1}^{k} w_k^i \left(\eta_i \varepsilon_i\right)^2 \\
\leqslant& \left(XA \left(\log X + H \log (Ak) \right)^2 + YB \left(\log Y + H \log (Bk) \right)^2 \right) k^{-2(\alpha_\eta + \alpha_\varepsilon) + \alpha_\eta + \alpha_\varepsilon} \\
=& \left(XA \left(\log X + H \log (Ak)^2 \right) + YB \left(\log Y + H \log (Bk) \right)^2 \right) k^{-\alpha_\eta - \alpha_\varepsilon} ,
\end{aligned}
$$

where the inequality follows from Lemma E.1. □

Lemma B.7 (Bounding **Term 4**).

$$\textbf{Term 4} \leqslant k^{\alpha_{\gamma_k}-2\alpha_\eta}(X+Y)\log\left(\frac{1}{\delta}\right) .$$

Proof.

$$\begin{aligned}
&\textbf{Term 4}\\
&=\sum_{i=1}^{k} w_k^i \eta_i^2 \left(X\underline{\tau}_i + Y\overline{\tau}_i\right)\\
&=\sum_{i=1}^{k} w_k^i \eta_i^2 \left(X\cdot\frac{1}{X}\sum_{h,x_h,a_h}\frac{1}{p_{1:h}^x(x_h)}\left(\frac{\mathbb{I}_h^k\left\{x_h,a_h\right\}}{\mu_{1:h}^k\left(x_h,a_h\right)+\gamma_k}-1\right)\right.\\
&\qquad\left.+Y\cdot\frac{1}{Y}\sum_{h,y_h,b_h}\frac{1}{p_{1:h}^y(y_h)}\left(\frac{\mathbb{I}_h^k\left\{y_h,b_h\right\}}{\nu_{1:h}^k\left(y_h,b_h\right)+\gamma_k}-1\right)\right)\\
&\leqslant \max_{1\leqslant i\leqslant k}\frac{w_k^i\eta_i^2(X+Y)}{\gamma_k}\log(\frac{1}{\delta})\\
&\leqslant k^{\alpha_\gamma-2\alpha_\eta}(X+Y)\log\left(\frac{1}{\delta}\right) ,
\end{aligned}$$

where the first inequality follows from that $\frac{1}{X}\frac{1}{p_{1:h}^x(x_h)} \leqslant 1$ for all (x_h,a_h) guaranteed by Lemma A.1 together with the use of Lemma B.15. □

Lemma B.8 (Bounding **Term 5**).

$$\textbf{Term 5} = \left(X^2A+Y^2B\right)k^{-\alpha_\eta+\alpha_\varepsilon} .$$

Proof.

$$\begin{aligned}
\textbf{Term 5} &= \sum_{i=1}^{k} w_k^i \eta_i^2 \left(X^2A+Y^2B\right)\\
&\leqslant \left(X^2A+Y^2B\right)k^{-2\alpha_\eta+\alpha_\eta+\alpha_\varepsilon}\\
&= \left(X^2A+Y^2B\right)k^{-\alpha_\eta+\alpha_\varepsilon} ,
\end{aligned}$$

where the inequality is given by Lemma E.1. □

Lemma B.9 (Bounding **Term 6**).

$$\textbf{Term 6} \leqslant (X+Y)^{\frac{1}{2}}\left(H\log\left(Ak\right)+H\log\left(Bk\right)\right)\cdot\log(k)k^{-\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}+\alpha_\eta+\alpha_\epsilon} .$$

Proof. To begin with, note that $\min_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A},h\in[H]}\mu_{1:h}^k(x_h,a_h)\geq\frac{1}{(Ak)^H}$ due to the definition of $\Pi_{\max}$ in Algorithm 1. Similarly, $\min_{(y_h,b_h)\in\mathcal{Y}_h\times\mathcal{B},h\in[H]}\nu_{1:h}^k(y_h,b_h)\geq\frac{1}{(Bk)^H}$ holds for the min-player. Further,

$$\begin{aligned}
&\textbf{Term 6}\\
&=\sum_{i=1}^{k} w_k^i\omega_i\\
&\leqslant (X+Y)^{\frac{1}{2}}\log\left(\frac{1}{(Ak)^H}\right)\log\left(\frac{1}{(Bk)^H}\right)\sum_{i=1}^{k}w_k^i i^{-\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}}\\
&\leqslant (X+Y)^{\frac{1}{2}}\log\left(\frac{1}{(Ak)^H}\right)\log\left(\frac{1}{(Bk)^H}\right)\log(k)k^{-\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}+\alpha_\eta+\alpha_\epsilon}\\
&\leqslant (X+Y)^{\frac{1}{2}}\left(H\log\left(Ak\right)+H\log\left(Bk\right)\right)\log(k)k^{-\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}+\alpha_\eta+\alpha_\epsilon} ,
\end{aligned}$$

where the first inequality is due to Lemma B.10 and the second inequality comes from Lemma E.1. □

B.3 Bounding the NE Gap of $(\mu^{k,\star}, \nu^{k,\star})$

Lemma B.10 (Bounding divergence difference)**.**

$$|w_k| = \mathcal{O}\left(\frac{(X+Y)^{\frac{1}{2}}\left(\ln\left((Ak)^H\right)+\ln\left((Bk)^H\right)\right)^2}{k^{\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}}}\right).$$

Proof. Again, note that $\min_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A},h\in[H]}\mu^k_{1:h}(x_h,a_h) \geq \frac{1}{(Ak)^H}$ and $\min_{(y_h,b_h)\in\mathcal{Y}_h\times\mathcal{B},h\in[H]}\nu^k_{1:h}(y_h,b_h) \geq \frac{1}{(Bk)^H}$. Therefore, it holds that

$$\begin{aligned}
&|w_k| \\
\leqslant&\left|D_\psi\left(\mu^{k+1,\star},\mu^{k+1}\right)-D_\psi\left(\mu^{k,\star},\mu^{k+1}\right)\right|+\left|D_\psi\left(\nu^{k+1,\star},\nu^{k-1}\right)-D_\psi\left(\nu^{k,\star},\nu^{k+1}\right)\right| \\
\leqslant&\left(\ln\left((Ak)^H\right)+\ln\left((Bk)^H\right)\right)\left(\left\|p^x\mu^{k+1,\star}-p^x\mu^{k,\star}\right\|_1+\left\|p^y\nu^{k+1,\star}-p^y\nu^{k,\star}\right\|_1\right) \\
\leqslant&\mathcal{O}\left(\frac{(X+Y)^{\frac{1}{2}}\left(\ln\left((Ak)^H\right)+\ln\left((Bk)^H\right)\right)^2}{k^{\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}}}\right),
\end{aligned}$$

where the second inequality is due to Lemma B.11 and the last inequality comes from Lemma B.12. □

Lemma B.11 (Bounding divergence using ℓ_1-norm)**.** $\forall\mu,\mu^1,\mu^2\in\Pi^k_{\max}$, *it holds that*

$$|D_\psi\left(\mu',\mu\right)-D_\psi\left(\mu^2,\mu\right)|\leqslant\mathcal{O}\left(\ln\left((Ak)^H\right)\left\|p^x\mu^1-p^x\mu^2\right\|_1\right).$$

Proof.

$$\begin{aligned}
&D_\psi\left(\mu',\mu\right)-D_\psi\left(\mu^2,\mu\right) \\
=&\sum_{h,(x_h,a_h)}p^x_{1:h}\left(x_h\right)\left(\mu^1_{1:h}\left(x_h,a_h\right)\log\frac{\mu^1_{1:h}\left(x_h,a_h\right)}{\mu_{1:h}\left(x_h\cdot a_h\right)}-\mu^2_{1:h}\left(x_h,a_h\right)\log\frac{\mu^2_{1:h}\left(x_h\cdot a_h\right)}{\mu_{1:h}\left(x_h,a_h\right)}\right) \\
=&\sum_{h,(x_h,a_h)}p^x_{1:h}\left(x_h\right)\left(\left(\mu^1_{1:h}\left(x_h,a_h\right)-\mu^2_{1:h}\left(x_h,a_h\right)\right)\log\frac{\mu^1_{1:h}\left(x_h,a_h\right)}{\mu_{1:h}\left(x_h\cdot a_h\right)}\right) \\
&+\sum_{h,(x_h,a_h)}p^x_{1:h}\left(x_h\right)\mu^1_{1:h}\left(x_h,a_h\right)\left(\log\frac{\mu^1_{1:h}\left(x_h,a_h\right)}{\mu_{1:h}\left(x_h\cdot a_h\right)}-\log\frac{\mu^2_{1:h}\left(x_h,a_h\right)}{\mu_{1:h}\left(x_h\cdot a_h\right)}\right) \\
\leqslant&\mathcal{O}\left(\ln\left((Ak)^H\right)\left\|p^x\mu^1-p^x\mu^2\right\|_1\right)-D_\psi(\mu^2,\mu^1) \\
\leqslant&\mathcal{O}\left(\ln\left((Ak)^H\right)\left\|p^x\mu^1-p^x\mu^2\right\|_1\right).
\end{aligned}$$

□

Lemma B.12 (Bounding ℓ_1-norm of the difference between $\mu^{k,\star}$ and $\mu^{k+1,\star}$)**.** *The ℓ_1-norm of the difference between $\mu^{k,\star}$ and $\mu^{k+1,\star}$ satisfies*

$$\left\|p^z\xi^{k+1,\star}-p^z\xi^{k,\star}\right\|_1=\mathcal{O}\left(\frac{(X+Y)^{\frac{1}{2}}\left(\ln\left((Ak)^H\right)+\ln\left((Bk)^H\right)\right)}{k^{\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}}}\right).$$

Proof. First note that, $\forall k,\forall(\mu,\nu)\in\Pi^k_{\max}\times\Pi^k_{\min}$, we have

$$\begin{aligned}
&f_k\left(\mu,\nu^{k,\star}\right)-f_k\left(\mu^{k,\star},\nu\right) \\
=&f_k\left(\mu,\nu^{k,\star}\right)-f_k\left(\mu^{k,\star},\nu^{k,\star}\right)+f_k\left(\mu^{k,\star},\nu^{k,\star}\right)-f_k\left(\mu^{k,\star},\nu\right) \\
\geqslant&f_k\left(\mu,\nu^{k,\star}\right)-f_k\left(\mu^{k,\star},\nu^{k,\star}\right)-\nabla_\mu f_k\left(\mu^{k,\star},\nu^{k,\star}\right)^\top\left(\mu-\mu^{k,\star}\right) \\
&-f_k\left(\mu^{k,\star},\nu\right)-\left(-f_k\left(\mu^{k,\star},\nu^{k,\star}\right)\right)-\left(-\nabla_\nu f_k\left(\mu^{k,\star},\nu^{k,\star}\right)^\top\left(\nu-\nu^{k,\star}\right)\right) \\
\geqslant&\varepsilon_k D_\psi\left(\mu,\mu^{k,\star}\right)+\varepsilon_k D_\psi\left(\nu,\nu^{k,\star}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \varepsilon_k \operatorname{KL}\left(p^x \mu, p^x \mu^{k,\star}\right)+\varepsilon_k \operatorname{KL}\left(p^y \nu, p^y \nu^{k,\star}\right) \\
&\geqslant \frac{1}{2} \varepsilon_k\left(\left\|p^x \mu-p^x \mu^{k,\star}\right\|_1^2+\left\|p^y \nu-p^y \nu^{k,\star}\right\|_1^2\right) \\
&\geqslant \frac{1}{4} \varepsilon_k\left\|p^z \xi-p^z \xi^{k,\star}\right\|_1^2 .
\end{aligned}
$$

Let $\mu^{k+1,\prime}=p_{k+1} \bar{\mu}+\left(1-p_{k+1}\right) \mu_{k+1}^{\star}$. Then $\forall h,\left(x_h, a_h\right)$,

$$
\mu^{k+1,\prime}\left(a_h \mid x_h\right) \geqslant p_{k+1} \frac{1}{A}+\left(1-p_{k+1}\right) \frac{1}{A(k+1)^2} \geqslant \frac{1}{A k^2},
$$

which means that $\mu^{k+1,\prime} \in \Pi_{\max }^k$. Similarly, we define $\nu^{k+1,\prime}$, which is such that $\nu^{k+1,\prime} \in \Pi_{\min }^k$. By previous analysis, we have

$$
f_k\left(\mu^{k+1,\prime}, \nu^{k,\star}\right)-f_k\left(\mu^{k,\star}, \nu^{k+1,\prime}\right) \geqslant \frac{1}{4} \varepsilon_k\left\|p^z \xi^{k+1,\prime}-p^z \xi^{k,\star}\right\|_1^2 . \tag{12}
$$

On the other hand, since $\left(\mu^{k,\star}, \nu^{k,\star}\right) \in \Pi_{\max }^{k+1} \times \Pi_{\min }^{k+1}$, we have

$$
f_{k+1}\left(\mu^{k,\star}, \nu^{k+1,\star}\right)-f_{k+1}\left(\mu^{k+1,\star}, \nu^{k,\star}\right) \geqslant \frac{1}{4} \varepsilon_{k+1}\left\|p^z \xi^{k,\star}-p^z \xi^{k+1,\star}\right\|_1^2 . \tag{13}
$$

Combing both sides, we have

$$
\begin{aligned}
& f_k\left(\mu^{k+1,\star}, \nu^{k,\star}\right)-f_k\left(\mu^{k,\star}, \nu^{k+1,\star}\right) \\
= & f_k\left(\mu^{k+1,\prime}, \nu^{k,\star}\right)-f_k\left(\mu^{k,\star}, \nu^{k+1,\prime}\right)+f_k\left(\mu^{k+1,\star}, \nu^{k,\star}\right)-f_k\left(\mu^{k+1,\prime}, \nu^{k,\star}\right) \\
& +f_k\left(\mu^{k,\star}, \nu^{k+1,\prime}\right)-f_k\left(\mu^{k,\star}, \nu^{k+1,\star}\right) \\
\geqslant & \frac{1}{4} \varepsilon_k\left\|p^z \xi^{k+1,\prime}-p^z \xi^{k,\star}\right\|_1^2+\left\langle\nabla_\mu f_k\left(\mu^{k+1,\prime}, \nu^{k,\star}\right), \mu^{k+1,\star}-\mu^{k+1,\star}\right\rangle \\
& +\left\langle\nabla_\nu f_k\left(\mu^{k,\star}, \nu^{k+1,\prime}\right), \nu^{k+1,\prime}-\nu^{k+1,\star}\right\rangle \\
\geqslant & \frac{1}{4} \varepsilon_k\left\|p^z \xi^{k+1,\prime}-p^z \xi^{k,\star}\right\|_1^2-\sup _{\mu \in \Pi_{\max }^{k+1}}\left\|\nabla_\mu f_k\left(\mu, \nu^{k,\star}\right)\right\|_{\infty}\left\|\mu^{k+1,\star}-\mu^{k+1,\prime}\right\|_1 \\
& -\sup _{\nu \in \Pi_{\min }^{k+1}}\left\|\nabla_\nu f_k\left(\mu^{k,\star}, \nu\right)\right\|_{\infty}\left\|\nu^{k+1,\prime}-\nu^{k+1,\star}\right\|_1 .
\end{aligned}
$$

Further using the fact that

$$
\begin{aligned}
& \left\|\nabla_\mu f_k\left(\mu, \nu^{k,\star}\right)\right\|_{\infty} \\
= & \max _{h,\left(x_h, a_h\right)}\left|\boldsymbol{G} \nu^{k,\star}\left[\left(x_h, a_h\right)\right]+\varepsilon_k p_{1: h}^x\left(x_h\right) \log \left[p^x \mu\right]\left[\left(x_h, a_h\right)\right]\right| \\
\leqslant & \max _{h,\left(x_h, a_h\right)}\left|\boldsymbol{G} \nu^{k,\star}\left[\left(x_h, a_h\right)\right]\right|+\left|\varepsilon_k p_{1: h}^x\left(x_h\right) \log \left[p^x \mu\right]\left[\left(x_h, a_h\right)\right]\right| \\
\leqslant & 1+k^{-\alpha \varepsilon}\left(\ln \left((A k)^H\right)+\ln \left((B k)^H\right)\right)=\mathcal{O}(1),
\end{aligned}
$$

and

$$
\begin{aligned}
\left\|\mu^{k+1,\star}-\mu^{k+1,\prime}\right\|_1 = & \left\|p_{k+1}\left(\bar{\mu}-\mu_{k+1}^{\star}\right)\right\|_1 \leqslant\left\|p_{k+1} \bar{\mu}\right\|_1+\left\|p_{k+1} \mu_{k+1}^{\star}\right\|_1 \\
\leqslant & p_{k+1} 2 X=\mathcal{O}\left(\frac{X+Y}{k^2}\right),
\end{aligned}
$$

one can deduce that

$$
\begin{aligned}
& f_k\left(\mu^{k+1,\star}, \nu^{k,\star}\right)-f_k\left(\mu^{k,\star}, \nu^{k+1,\star}\right) \\
\geqslant & \frac{1}{8} \varepsilon_k\left\|p^z \xi^{k+1,\star}-p^z \xi^{k,\star}\right\|_1^2-\frac{1}{4} \varepsilon_k\left\|p^z \xi^{k+1,\star}-p^z \xi^{k+1,\star}\right\|_1^2-\mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
\geqslant & \frac{1}{8} \varepsilon_k\left\|p^z \xi^{k+1,\star}-p^z \xi^{k,\star}\right\|_1^2-\frac{1}{4} \varepsilon_k\left(2\left(\left\|p^x \mu^{k+1,\prime}-p^x \mu^{k+1,\star}\right\|_1^2+\left\|p^y \nu^{k+1,\prime}-p^y \nu^{k+1,\star}\right\|_1^2\right)\right) \\
& -\mathcal{O}\left(\frac{X+Y}{k^3}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\geqslant \frac{1}{8}\varepsilon_k \left\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\right\|_1^2 - \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&\quad - \frac{1}{4}\varepsilon_k \left(4\left(\|p_{k+1}p^x\bar{\mu}\|_1^2 + \left\|p_{k+1}p^x\mu^{k+1,\star}\right\|_1^2\right) + 4\left(\|p_{k+1}p^y\bar{\nu}\|_1^2 + \left\|p_{k+1}p^y\nu^{k+1,\star}\right\|_1^2\right)\right) \\
&= \frac{1}{8}\varepsilon_k \left\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\right\|_1^2 - \mathcal{O}\left(\frac{X+Y}{k^3}\right) - \mathcal{O}\left(\frac{1}{k^6}\right) \\
&\geqslant \frac{1}{8}\varepsilon_{k+1} \left\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\right\|_1^2 - \mathcal{O}\left(\frac{X+Y}{k^3}\right).
\end{aligned}
$$

Combining with Eq. (13), we have

$$
\begin{aligned}
&\frac{3}{8}\varepsilon_{k+1}\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\|_1^2 \\
&\leqslant f_{k+1}\left(\mu^{k,\star},\nu^{k+1,\star}\right) - f_k\left(\mu^{k,\star},\nu^{k+1,\star}\right) - f_{k+1}\left(\mu^{k+1,\star},\nu^{k,\star}\right) + f_k\left(\mu^{k+1,\star},\nu^{k,\star}\right) + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&= \bar{f}_k\left(\mu^{k,\star},\nu^{k+1,\star}\right) - \bar{f}_k\left(\mu^{k+1,\star},\nu^{k,\star}\right) + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \quad \left(\bar{f}_k(\mu,\nu) := f_{k+1}(\mu,\nu) - f_k(\mu,\nu)\right) \\
&= \bar{f}_k\left(\mu^{k,\star},\nu^{k+1,\star}\right) - \bar{f}_k\left(\mu^{k+1,\star},\nu^{k+1,\star}\right) + \bar{f}_k\left(\mu^{k+1,\star},\nu^{k+1,\star}\right) - \bar{f}_k\left(\mu^{k+1,\star},\nu^{k,k}\right) + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&\leqslant \left\langle \nabla_\mu \bar{f}_k\left(\mu^{k,\star},\nu^{k+1,\star}\right), \mu^{k,\star} - \mu^{k+1,\star}\right\rangle + \left\langle \nabla_\nu \bar{f}_k\left(\mu^{k+1,\star},\nu^{k,\star}\right), \nu^{k+1,\star} - \nu^{k,\star}\right\rangle + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&= \left\langle \nabla_\mu \bar{f}_k\left(\mu^{k,\star},\nu^{k+1,\star}\right)/p^x, p^x\left(\mu^{k,\star} - \mu^{k+1,\star}\right)\right\rangle + \left\langle \nabla_\nu \bar{f}_k\left(\mu^{k+1,\star},\nu^{k,\star}\right)/p^y, p^y\left(\nu^{k+1,\star} - \nu^{k,\star}\right)\right\rangle + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&\leqslant \left\|\nabla_\mu \bar{f}_k\left(\mu^{k,\star},\nu^{k+1,\star}\right)/p^x\right\|_\infty \left\|p^x\left(\mu^{k,\star} - \mu^{k+1,\star}\right)\right\|_1 + \left\|\nabla_\nu \bar{f}_k\left(\mu^{k+1,\star},\nu^{k,\star}\right)/p^y\right\|_\infty \left\|p^y\left(\nu^{k+1,\star} - \nu^{k,\star}\right)\right\|_1 \\
&\quad + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&\leqslant \left(\sup_{\mu\in\Pi_{\max}^k} \left\|\nabla_\mu \bar{f}_k\left(\mu,\nu^{k+1,\star}\right)/p^x\right\|_\infty + \sup_{\nu\in\Pi_{\min}^k} \left\|\nabla_\nu \bar{f}_k\left(\mu^{k+1,\star},\nu\right)/p^y\right\|_\infty\right) \left\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\right\|_1 + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&\leqslant \left(\sup_{\mu\in\Pi_{\max}^k} \max_{h,(x_h,a_h)} \left|(\varepsilon_k - \varepsilon_{k+1})\log[p^x\mu][(x_h,a_h)]\right| + \sup_{\nu\in\Pi_{\min}^k} \max_{h,(y_h,b_h)} \left|(\varepsilon_k - \varepsilon_{k+1})\log[p^y\nu][(y_h,b_h)]\right|\right) \\
&\quad \cdot \left\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\right\|_1 + \mathcal{O}\left(\frac{X+Y}{k^3}\right) \\
&= \mathcal{O}\left((\varepsilon_k - \varepsilon_{k+1})\left(\ln\left((Ak)^H\right) + \ln\left((Bk)^H\right)\right)\left\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\right\|_1 + \frac{X+Y}{k^3}\right).
\end{aligned}
$$

In what follows, we slightly abuse the notations by denoting $m_k = (Ak)^H(Bk)^H$. Solving the above equation leads to

$$
\begin{aligned}
&\left\|p^z\xi^{k+1,\star} - p^z\xi^{k,\star}\right\|_1 \\
&\leqslant \frac{(\varepsilon_k - \varepsilon_{k+1})\log(m_k) + \sqrt{(\varepsilon_k - \varepsilon_{k+1})^2\log^2(m_k) + \varepsilon_{k+1}\frac{X+Y}{k^3}}}{\varepsilon_{k+1}} \\
&\leqslant \frac{(\varepsilon_k - \varepsilon_{k+1})}{\varepsilon_{k+1}}\log(m_k) + \sqrt{\frac{X+Y}{\varepsilon_{k+1}k^3}} \\
&\leqslant \frac{\log(m_k)}{k} + \sqrt{\frac{X+Y}{\varepsilon_{k+1}k^3}} = \mathcal{O}\left(\frac{(X+Y)^{\frac{1}{2}}\log(m_k)}{k^{\min\left\{1,\frac{3}{2}-\frac{\alpha}{2}\right\}}}\right).
\end{aligned}
$$

In the last inequality of the above display, we use the fact that

$$
\frac{(\varepsilon_k - \varepsilon_{k+1})}{\varepsilon_{k+1}} = \frac{k^{-\alpha_\epsilon}}{(k+1)^{-\alpha_\epsilon}} = (1+\frac{1}{k})^{\alpha_\epsilon} - 1 = \mathcal{O}\left(\frac{\alpha_\epsilon}{k}\right),
$$

by Taylor expansion. □

Lemma B.13. *Let $\{c_i\}_{i=1}^k$ be fixed positive numbers. Then with probability at least $1-\delta$, it holds that*

$$\sum_{i=1}^k c_i \left\langle \mu^i, \ell^{i,x} - \hat{\ell}^{i,x} \right\rangle \leqslant XA \sum_{i=1}^k c_i \gamma_{k_i} + H \sqrt{2 \sum_{i=1}^k c_i^2 \log \frac{1}{\delta}} \,.$$

Proof. To begin with, notice that

$$\sum_{i=1}^k c_i \left\langle \mu^i, \ell^{i,x} - \hat{\ell}^{i,x} \right\rangle = \sum_{i=1}^k c_i \left\langle \mu^i, \ell^{i,x} - \mathbb{E}_{i-1}\left[\hat{\ell}^{i,x}\right] \right\rangle + \sum_{i=1}^k c_i \left\langle \mu^i, \mathbb{E}_{i-1}\left[\hat{\ell}^{i,x}\right] - \hat{\ell}^{i,x} \right\rangle \,.$$

For the first part, we have

$$\begin{aligned}
&\sum_{i=1}^k c_i \left\langle \mu^i, \ell^{i,x} - \mathbb{E}_{i-1}\left[\hat{\ell}^{i,x}\right] \right\rangle \\
=&\sum_{i=1}^k c_i \sum_{h,x_h,a_h} \mu^i_{1:h}\left(x_h, a_h\right) \ell^{i,x}_{[(x_h,a_h)]} \left(1 - \frac{\mu^i_{1:h}\left(x_h,a_h\right)}{\mu^i_{1:h}\left(x_h,a_h\right) + \gamma_{k_i}}\right) \\
\leqslant&\sum_{i=1}^k c_i \gamma_{k_i} \sum_{h,x_h,a_h} \ell^{i,x}_{[(x_h,a_h)]} \\
\leqslant&\sum_{i=1}^k c_i \gamma_{k_i} XA \,,
\end{aligned}$$

where the last inequality comes from $\ell[(x_h,a_h)]^{i,x} \leqslant 1$ for all $(x_h,a_h) \in \mathcal{X} \times \mathcal{A}$.

For the second part, taking $\delta = \exp\left(\frac{-\varepsilon^2}{2\sum_{i=1}^k c_i^2 H^2}\right)$, $\varepsilon = \sqrt{2\sum_{i=1}^k c_i^2 H^2 \log\left(\frac{1}{\delta}\right)}$ and using Azuma-Hoeffding inequality, it holds with probability at least $1-\delta$ that

$$\sum_{i=1}^k c_i \left\langle \mu^i, \mathbb{E}_{i-1}\left[\hat{\ell}^{i,x}\right] - \hat{\ell}^{i,x} \right\rangle \leqslant \sqrt{2 \sum_{i=1}^k c_i^2 H^2 \log\left(\frac{1}{\delta}\right)} \,.$$

The proof is concluded by combining the upper bounds of the two parts above. ☐

Lemma B.14. *Let $\delta \in (0,1)$ and $\{\gamma_{k_i}\}_{i=1}^k \in (0,+\infty)^k$. Fix $h \in [H]$. For any coefficient sequence $\left\{c^i\right\}_{i=1}^k$ s.t. $c^i \in \left[0, 2{\gamma_k}^i\right]^{XA}$ is $\mathcal{F}_{i-1}$ - measurable, with probability $1-\delta$, we have*

$$\sum_{i=r}^k w_i \left\langle c_i, \hat{\ell}_i - \ell_i \right\rangle \leqslant \max_{1\leqslant i\leqslant k} w_i \log \frac{1}{\delta} \,.$$

Proof. Define $w = \max_{1\leqslant i\leqslant k} w_i$. Hence

$$\begin{aligned}
&w^i \hat{\ell}^i\left(x_h, a_h\right) \\
=&\frac{w_i \mathbb{I}_{i,h}\left\{x_h,a_h\right\}\left(1 - r_h^i\right)}{\mu^i_{1:h}\left(x_h,a_h\right) + r_i} \\
\leqslant&\frac{w_i \mathbb{I}_{i,h}\left\{x_h,a_h\right\}\left(1 - r_h^i\right)}{\mu^i_{i,h}\left(x_h,a_h\right) + r_i \frac{w_i\left(1-r_h^i\right)\mathbb{I}_{i,h}\left\{x_h,a_h\right\}}{w}} \\
=&\frac{w}{2\gamma_{k_i}} \frac{\frac{2\gamma_{k_i} w_i\left(1-r_h^i\right)\mathbb{I}_{i,h}\left\{x_h,a_h\right\}}{w\mu^i_{i:h}\left(x_h,a_h\right)}}{1 + \frac{\gamma_{k_i} w_i\left(1-r_h^i\right)\mathbb{I}_{i,h}\left\{x_h,a_h\right\}}{w\mu^i_{i:h}\left(x_h,a_h\right)}}
\end{aligned}$$

$$\leqslant \frac{w}{2\gamma_{k_i}} \log\left(1 + \frac{2\gamma_{k_i} w_i \left(1 - r_h^i\right) \mathbb{I}_{i,h}\left\{x_h, a_h\right\}}{w\mu_{i:h}^i\left(x_h, a_h\right)}\right) .$$

Denote by $\hat{S}_h^i = \frac{w_i}{w}\left\langle c^i, \hat{\ell}_h^i \right\rangle$, $S_h^i = \frac{w_i}{w}\left\langle c^i, \ell_h^i \right\rangle$. Then

$$\begin{aligned}
&\mathbb{E}_{i-1}[\exp(\hat{S}^i)] \\
\leqslant &\mathbb{E}_{i-1}\left[\exp\left(\sum_{(x_h,a_h)\in\mathcal{X}\times\mathcal{A}} \frac{c^i(x_h,a_h)}{2\gamma_{k_i}} \log\left(1 + \frac{2\gamma_{k_i} w_i \left(1 - r_h^i\right) \mathbb{I}_{i,h}\left\{x_h, a_h\right\}}{w\mu_{i:h}^i\left(x_h, a_h\right)}\right)\right)\right] \\
\leqslant &\mathbb{E}_{i-1}\left[\prod_{(x_h,a_h)\in\mathcal{X}\times\mathcal{A}} \left(1 + \frac{c_i(x_h,a_h) w_i \left(1 - r_h^i\right) \mathbb{I}_{i,h}\left\{x_h, a_h\right\}}{w\mu_{i:h}^i\left(x_h, a_h\right)}\right)\right] \\
= &\mathbb{E}_{i-1}\left[1 + \sum_{(x_h,a_h)\in\mathcal{X}\times\mathcal{A}} \frac{c^i\left(x_h, a_h\right) w_i \left(1 - r_h^i\right) \mathbb{I}_{i,h}\left\{x_h, a_h\right\}}{w\mu_{i:h}^i\left(x_h, a_h\right)}\right] \\
= &1 + S_h^i \leqslant \exp\left(S_h^i\right) .
\end{aligned}$$

Finally, one can see that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^k \left(\hat{S}_h^i - S_h^i\right) \geqslant \log\frac{1}{\delta}\right] \\
= &\mathbb{E}\left[\exp\left(\sum_{i=1}^k \left(\hat{S}_h^i - S_h^i\right)\right) \geqslant \frac{1}{\delta}\right] \\
\leqslant &\delta\mathbb{E}\left[\exp\left(\sum_{i=1}^k \left(\hat{S}_h^i - S_h^i\right)\right)\right] \\
= &\delta\mathbb{E}\left[\left[\mathbb{E}_{k-1}\left[\exp\left(\sum_{i=1}^k \left(\hat{S}_h^i - S_h^i\right)\right)\right]\right]\right] \\
= &\delta\mathbb{E}\left[\exp\left(\sum_{i=1}^{k-1} \left(\hat{S}_h^i - S_h^i\right)\right) \left[\mathbb{E}_{k-1}\left[\exp\left(\hat{S}_h^k - S_h^k\right)\right]\right] \leqslant \ldots \leqslant \delta\right. .
\end{aligned}$$

□

Lemma B.15. *Let $\{c_i\}_{i=1}^k$ be fixed positive numbers. Fix $h \in [H]$. Then $\forall$ sequence $\{q_i\}_{i=1}^k \in [0,1]^{XA}$ s.t. q^i is $\mathcal{F}_{i-1}$ - measurable, with probability at least $1-\delta$,*

$$\sum_{i=1}^k c_i \left\langle q_i, \hat{\ell}_h^i - \ell_h^i \right\rangle \leqslant \max_{1\leqslant i\leqslant k} \frac{c_i}{\gamma_{k_i}} \log\left(\frac{1}{\delta}\right) .$$

Proof. Noticing that $\{\gamma_{k_i}\}_{i=1}^k$ is decreasing and $\|q^i\|_\infty \leqslant 1$, applying Lemma B.14, we arrive at

$$\sum_{i=1}^k c_i \left\langle q^i, \hat{\ell}_h^i - \ell_h^i \right\rangle = \sum_{i=1}^k \frac{c_i}{2\gamma_{k_i}} \left\langle 2\gamma_{k_i} q^i, \hat{\ell}_h^i - \ell_h^i \right\rangle \leqslant \max_{1\leqslant i\leqslant k} \frac{c_i}{\gamma_{k_i}} \log\left(\frac{1}{\delta}\right) .$$

□

B.4 Proof of Theorem 5.1

We are now ready to present the proof of our main result.

Proof of Theorem 5.1. Putting Lemma B.3, B.4, B.5, B.6, B.7, B.8, B.9 together, we have

$$
\begin{aligned}
& D_\psi\left(\xi^{k+1,\star}, \xi^{k+1}\right) \\
=& \mathcal{O}\Bigg((XA+YB)\ln(k)k^{-\alpha_{\gamma_k}+\alpha_\varepsilon}+k^{-\frac{\alpha_\eta}{2}+\frac{\alpha_\varepsilon}{2}}\log\left(\frac{k^2}{\delta}\right)+k^{-\alpha_\eta+\alpha_{\gamma_k}}\log\left(\frac{k^2}{\delta}\right) \\
& +\left(XA\left(\log X+H\log(Ak)^2+YB\left(\log Y+H\log(Bk)\right)^2\right)k^{-\alpha_\eta-\alpha_\varepsilon}\right. \\
& +k^{\alpha_{\gamma_k}-2\alpha_\eta}(X+Y)\log\left(\frac{1}{\delta}\right)+(X^2A+Y^2B)k^{-\alpha_\eta+\alpha_\epsilon} \\
& +(X+Y)^{\frac{1}{2}}\left(H\log(Ak)+H\log(Bk)\right)\left(\log X+H\log(k)+\log Y+H\log(Bk)\right) \\
& \quad\cdot\log(k)k^{-\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}+\alpha_\eta+\alpha_\epsilon} \\
=& \mathcal{O}\Bigg(\Big[k^{-\frac{1}{4}}(XA+YB)+k^{-\frac{1}{4}}+k^{-\frac{1}{4}}+(XA+YB)H^2k^{-\frac{3}{4}}+(X+Y)k^{-\frac{7}{8}}+\left(X^2A+Y^2B\right)k^{-\frac{1}{2}} \\
& +(X+Y)^{\frac{1}{2}}H^2K^{-\frac{1}{4}}\Big]\left(\log^2\left(XAk/\delta\right)+\log^2\left(YBk/\delta\right)\right)\log(K)\Bigg)\ .
\end{aligned}
$$

Moreover, note that

$$
\begin{aligned}
& \operatorname{NEGap}\left(\xi^k\right) \\
=& \sup_{\mu\in\Pi_{\max},\nu\in\Pi_{\min}} f\left(\mu^k,\nu\right)-f\left(\mu,\nu^k\right) \\
=& f\left(\mu^{k,\star},\nu\right)-f\left(\mu^{k,\star},\nu\right)+f\left(\mu^k,\nu\right)-f\left(\mu,\nu^k\right)+f\left(\mu,\nu^{k,\star}\right)-f\left(\mu,\nu^{k,\star}\right) \\
\leqslant& \operatorname{NEGap}\left(\xi^{k,\star}\right)+\left(\mu^k-\mu^{k,\star}\right)^\top \boldsymbol{G}\nu+\mu^\top\boldsymbol{G}\left(\nu^{k,\star}-\nu^k\right) \\
\leqslant& \operatorname{NEGap}\left(\xi^{k,\star}\right)+\left\langle p^x\left(\mu^k-\mu^{k,\star}\right),\boldsymbol{G}\nu/p^x\right\rangle+\left\langle p^y\left(\nu^k-\nu^{k,\star}\right),\boldsymbol{G}^\top\mu/p^y\right\rangle \\
\leqslant& \operatorname{NEGap}\left(\xi^{k,\star}\right)+\left\|p^x\left(\mu^k-\mu^{k,\star}\right)\right\|_1\|\boldsymbol{G}\nu/p^x\|_\infty+\left\|p^y\left(\nu^k-\nu^{k,\star}\right)\right\|_1\|\boldsymbol{G}^\top\mu/p^y\|_\infty \\
\leqslant& \operatorname{NEGap}\left(\xi^{k,\star}\right)+X\left\|p^x\left(\mu^k-\mu^{k,\star}\right)\right\|_1+Y\left\|p^y\left(\nu^k-\nu^{k,\star}\right)\right\|_1 \\
\leqslant& \varepsilon_k H\left(\ln(XA)+\ln(YB)\right)+\mathcal{O}\left(\frac{XAH}{k}+\frac{YBH}{k}\right) \\
& +\mathcal{O}\left(X\sqrt{\operatorname{KL}\left(p^x\mu^{k,\star},p^x\mu^k\right)}+Y\sqrt{\operatorname{KL}\left(p^y\nu^{k,\star},p^y\nu^k\right)}\right) \\
\leqslant& \varepsilon_k H\left(\ln(XA)+\ln(YB)\right)+\mathcal{O}\left(\frac{XAH}{k}+\frac{YBH}{k}\right) \\
& +\mathcal{O}\left((X+Y)\sqrt{\operatorname{KL}\left(p^x\mu^{k,\star},p^x\mu^k\right)+\operatorname{KL}\left(p^y\nu^{k,\star},p^y\nu^k\right)}\right) \\
\leqslant& \varepsilon_k H\left(\ln(XA)+\ln(YB)\right)+\mathcal{O}\left(\frac{XAH}{k}+\frac{YBH}{k}\right)+\mathcal{O}\left((X+Y)\sqrt{\operatorname{KL}\left(p^z\xi^{k,\star},p^z\xi^k\right)}\right) \\
\leqslant& \varepsilon_k H\left(\ln(XA)+\ln(YB)\right)+\mathcal{O}\left(\frac{XAH}{k}+\frac{YBH}{k}\right)+\mathcal{O}\left((X+Y)\sqrt{D_\psi\left(\xi^{k,\star},\xi^k\right)}\right)\ ,
\end{aligned}
$$

where $\boldsymbol{G}\nu/p^x\in\mathbb{R}^{XA}$ is defined as $(\boldsymbol{G}\nu/p^x)[(x_h,a_h)]=(\boldsymbol{G}\nu)[(x_h,a_h)]/p^x_{1:h}(x_h)$ and similarly for $\boldsymbol{G}^\top\mu/p^y$.

Therefore, we can see that

$$
\begin{aligned}
& \operatorname{NEGap}(\mu^k,\nu^k) \\
=& \mathcal{O}\Bigg((X+Y)\left[k^{-\frac{1}{8}}(XA+YB)^{\frac{1}{2}}+(XA+YB)^{\frac{1}{2}}Hk^{-\frac{3}{8}}+\left(X^2A+Y^2B\right)^{\frac{1}{2}}k^{-\frac{1}{4}}+(X+Y)^{\frac{1}{4}}HK^{-\frac{1}{8}}\right] \\
& \cdot\left(\log\left(XAk/\delta\right)+\log\left(YBk/\delta\right)\right)\log^{\frac{1}{2}}(k)+k^{-\frac{1}{8}}H(\ln(XA)+\ln(YB))+\frac{XAB}{k}+\frac{YBH}{k}\Bigg) \\
=& \widetilde{\mathcal{O}}\Bigg((X+Y)\left[k^{-\frac{1}{8}}(XA+YB)^{\frac{1}{2}}+(XA+YB)^{\frac{1}{2}}HK^{-\frac{3}{8}}+\left(X^2A+Y^2B\right)^{\frac{1}{2}}k^{-\frac{1}{4}}+(X+Y)^{\frac{1}{4}}Hk^{-\frac{1}{8}}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{(XAH+YBH)}{k}\Bigg) \\
=&\widetilde{\mathcal{O}}\left((X+Y)k^{-\frac{1}{8}}\left[(XA+YB)^{\frac{1}{2}}+(X+Y)^{\frac{1}{4}}H\right]\right) ,
\end{aligned}
$$

where the last equality holds when $k \geqslant \max\{H^4, {(X^2A+Y^2B)^4}/{(XA+YB)^4}, {(XA+YB)^{8/7}}/{(X+Y)^{10/7}}\}$. □

C LAST-ITERATE CONVERGENCE RATE IN EXPECTATION

Theorem C.1. *With the same condition as in Theorem 5.1, Algorithm 1 guarantees that*

$$
\mathbb{E}\left[\mathrm{NEGap}(\mu^k,\nu^k)\right]=\widetilde{\mathcal{O}}\left(\left((X+Y)^{\frac{1}{4}}H+\sqrt{(X^2A+Y^2B)}\right)k^{-\frac{1}{6}}\right) .
$$

Proof. With the same arguments as in the proof of Theorem 5.1, we have

$$
\begin{aligned}
&D_\psi\left(\xi^{k+1,x},\xi^{k+1}\right) \\
\leqslant&(1-\eta_k\varepsilon_k)D_\psi\left(\xi^{k,\star},\xi^k\right)+\eta_k^2\left(X\underline{\tau}_k+Y\overline{\tau}_k\right)+\eta_k^2\left(X^2A+Y^2B\right)+\eta_k\rho_k+\eta_k\sigma_k+\omega_k \\
&+\eta_k^2XA\varepsilon_k^2\left(\log X+H\log(Ak)\right)^2+\eta_k^2YB\varepsilon_k^2\left(\log Y+H\log(Bk)\right)^2 .
\end{aligned}
$$

Taking conditional expectation $\mathbb{E}_{k-1}[\cdot]$ on both sides and by noticing the fact that $\mathbb{E}_{k-1}\left[\tau_k\right]<0$, $\mathbb{E}_{k-1}\left[\rho_k\right]=0$, and $\mathbb{E}_{k-1}\left[\sigma_k\right]=0$, we have

$$
\begin{aligned}
&\mathbb{E}_{k-1}\left[D_\psi\left(\xi^{k+1,x},\xi^{k+1}\right)\right] \\
\leqslant&(1-\eta_k\varepsilon_k)D_\psi\left(\xi^{k,\star},\xi^k\right)+\eta_k^2\left(X^2A+Y^2B\right)+\mathbb{E}_{k-1}\left[\omega_k\right] \\
&+\eta_k^2XA\varepsilon_k^2\left(\log X+H\log(Ak)\right)^2+\eta_k^2YB\varepsilon_k^2\left(\log Y+H\log(Bk)\right)^2 .
\end{aligned}
$$

Expanding the recursion in the above display leads to

$$
\begin{aligned}
&\mathbb{E}\left[D_\psi\left(\xi^{k+1,\star},\xi^{k+1}\right)\right] \\
\leqslant&\mathbb{E}\left[\sum_{i=1}^k w_k^i\omega_i\right]+XA\left(\log X+H\log(Ak)\right)^2\sum_{i=1}^k w_k^i\left(\eta_i\varepsilon_i\right)^2+YB\left(\log Y+H\log(Bk)\right)^2\sum_{i=1}^k w_k^i\left(\eta_i\varepsilon_i\right)^2 \\
&+\sum_{i=1}^k w_k^i\eta_i^2\left(X^2A+Y^2B\right) \\
\leqslant&(X+Y)^{\frac{1}{2}}(H\log(Ak)+H\log(Bk))(\log X+H\log(Ak)+\log Y+H\log(Bk)) \\
&\cdot\log(k)k^{-\min\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\}-\alpha_\eta+\alpha_\varepsilon} \\
&+\left(XA\left(\log X+H\log(Ak)\right)^2+YB\left(\log Y+H\log(Bk)\right)^2\right)k^{-\alpha_\eta-\alpha_\varepsilon}+\left(X^2A+Y^2B\right)k^{-\alpha_\eta+\alpha_\varepsilon} \\
=&\widetilde{\mathcal{O}}\left((X+Y)^{\frac{1}{2}}H^2k^{-\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}+\alpha_\eta+\alpha_\varepsilon}+(XA+YB)H^2k^{-\alpha_\eta-\alpha_\varepsilon}+\left(X^2A+Y^2B\right)k^{-\alpha_\eta+\alpha_\varepsilon}\right) .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathrm{NEGap}(\mu^k,\nu^k) \\
=&\widetilde{\mathcal{O}}\Bigg(\varepsilon_kH+\frac{XAH}{k}+\frac{YBH}{k} \\
&+(X+Y)\left[(X+Y)^{\frac{1}{4}}Hk^{\left(-\min\left\{1,\frac{3}{2}-\frac{\alpha_\varepsilon}{2}\right\}+\alpha_\eta+\alpha_\varepsilon\right)/2}+\sqrt{(XA+YB)}+Hk^{\frac{-\alpha_\eta-\alpha_\varepsilon}{2}}\right. \\
&\left.+\sqrt{(X^2A+Y^2B)}k^{\frac{-\alpha_\eta+\alpha_\varepsilon}{2}}\right]\Bigg) \\
=&\widetilde{\mathcal{O}}\Bigg(k^{-\frac{1}{6}}H+\frac{XAH}{k}+\frac{YBH}{k}+(X+Y)\left[(X+Y)^{\frac{1}{4}}Hk^{-\frac{1}{6}}+\sqrt{(XA+YB)}Hk^{-\frac{1}{3}}\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\ \sqrt{X^2A+Y^2B}k^{-\frac{1}{6}}\Big]\Big) \\
=&\widetilde{\mathcal{O}}\left((X+Y)\left[(X+Y)^{\frac{1}{4}}H\ +\ \sqrt{(X^2A+Y^2B)}\right]k^{-\frac{1}{6}}\right)\,.
\end{aligned}
$$

□

D Proof of Lower Bound of Last-iterate Convergence

Proof of Theorem 5.3. Let $\mathrm{NEGap}_k := \mathrm{NEGap}\left(\mu^k,\nu^k\right)$ with $\left(\mu^k,\nu^k\right)$ as the policy profile generated by some algorithm Alg. Suppose that Alg leans the IIEFG with the last-iterate convergence rate of $\mathrm{NEGap}_k = \Theta\left(f(X,A)k^{-\alpha}\right)$ for some $\alpha\in(0,1)$, where $f^{\mathrm{Alg}}(X,A)$ denotes the polynomial dependence on X and A of NEGap_k.

Fix some $K\geqslant\max(XA,YB)$. Consider the regret defined as follows (Kozuno et al., 2021; Bai et al., 2022; Fiegel et al., 2023):

$$
\mathrm{Reg}_K(\mathrm{Alg}) = \sup_{\mu\in\Pi_{\max}}\sum_{k=1}^{K}\left\langle\mu^k-\mu,\boldsymbol{G}\nu^k\right\rangle\,,
$$

where $\{\nu^k\}_{k\in[K]}$ is potentially generated by an adversary. Then, one can deduce that

$$
\begin{aligned}
\mathrm{Reg}_K(\mathrm{Alg}) = &\sup_{\mu\in\Pi_{\max}}\sum_{k=1}^{K}\langle\mu_k-\mu,\boldsymbol{G}\nu_k\rangle \qquad (14)\\
\leqslant&\sum_{k=1}^{K}\sup_{\mu\in\Pi_{\max}}\ \langle\mu_k-\mu,\boldsymbol{G}\nu_k\rangle\\
=&\sum_{k=1}^{K}\sup\nolimits_{\mu\in\Pi_{\max}}\mu_k^\top\boldsymbol{G}\nu_k-\mu^\top\boldsymbol{G}\nu_k\\
\leqslant&\sum_{k=1}^{K}\sup\nolimits_{\mu\in\Pi_{\max},\nu\in\Pi_{\min}}\mu_k^\top\boldsymbol{G}\nu-\mu^\top\boldsymbol{G}\nu_k\\
=&\sum_{k=1}^{K}\mathrm{NEGap}_k\\
=&\Theta\left(f(X,A)\sum_{k=1}^{K}k^{-\alpha}\right)\\
=&\Theta\left(f(X,A)K^{1-\alpha}\right)\,. \qquad (15)
\end{aligned}
$$

On the other hand, by Theorem 6 of Bai et al. (2022) (see also Theorem 3.1 fo Fiegel et al. (2023)), we have

$$
\mathrm{Reg}_K(\mathrm{Alg})\geqslant\Omega(\sqrt{AXK})\,. \qquad (16)
$$

Combining Eq. (14) and Eq. (16), we have

$$
\Omega(\sqrt{AXK})\leqslant\Theta\left(f(X,A)K^{1-\alpha}\right)\,.
$$

We now further consider the following three cases:

- If $\alpha>\frac{1}{2}$, then $\sqrt{AX}\leqslant f(X,A)K^{\frac{1}{2}-\alpha}$. However, this does not hold for any f, when K is large enough;
- If $\alpha=\frac{1}{2}$, it must hold that $\sqrt{AX}\leqslant f(X,A)$;
- If $\alpha<\frac{1}{2}$, then $\sqrt{AX}\leqslant f(X,A)K^{\frac{1}{2}-\alpha}$. This holds for all f, including $f(X,A)=1$ when K is large enough. In this case, the "minimal" f is $f(X,A)=1$, implying that the minimal possible convergence rate of NEGap_k in this case is $\mathrm{NEGap}_k=\Theta\left(k^{-\alpha}\right)$.

Taking the above three cases into account, the minimal possible convergence rate is

$$\begin{aligned}&\min\left\{\Theta\left(\sqrt{XA}k^{-\frac{1}{2}}\right),\Theta\left(k^{-\alpha}\right)\right\}\quad(\alpha>\frac{1}{2})\\=&\Theta\left(\sqrt{XA}k^{-\frac{1}{2}}\right).\end{aligned}$$

Analogously, we can prove that $\mathrm{NEGap}_k \geq \Theta(\sqrt{YB}k^{-\frac{1}{2}})$. Therefore, we have

$$\mathrm{NEGap}_k \geq \Theta\left(\left(\sqrt{XA}+\sqrt{YB}\right)k^{-\frac{1}{2}}\right).$$

The proof is concluded by noticing that the above holds for all algorithms. □

E AUXILIARY LEMMAS

Lemma E.1 (Lemma 1 of Cai et al. (2023))**.** *Let* $0 < h < 1, 0 \leqslant k \leqslant 2$*, and let* $t \geqslant \left(\frac{24}{1-h}\ln\frac{12}{1-h}\right)^{\frac{1}{1-h}}$*. Then*

$$\sum_{i=1}^{t}\left(i^{-k}\prod_{j=i+1}^{t}\left(1-j^{-h}\right)\right)\leqslant 9\ln(t)t^{-k+h}.$$

Lemma E.2 (Lemma 2 of Cai et al. (2023))**.** *Let* $0 < h < 1, 0 \leqslant k \leqslant 2$*, and let* $t \geqslant \left(\frac{24}{1-h}\ln\frac{12}{1-h}\right)^{\frac{1}{1-h}}$*. Then*

$$\max_{1\leqslant i\leqslant t}\left(i^{-k}\prod_{j=i+1}^{t}\left(1-j^{-h}\right)\right)\leqslant 4t^{-k}.$$

Lemma E.3 (Lemma 20 of Bai et al. (2020))**.** *Let* $c_1, c_2, \ldots, c_t$ *be fixed positive numbers. Then with probability at least* $1-\delta$*,*

$$\sum_{i=1}^{t}c_i\left\langle x_i,\ell_i-\widehat{\ell}_i\right\rangle=\mathcal{O}\left(A\sum_{i=1}^{t}\beta_i c_i+\sqrt{\ln(A/\delta)\sum_{i=1}^{t}c_i^2}\right).$$

F OPTIMIZATION PROBLEM IN EQ. (3)

Algorithm 3 Frank-Wolfe-type Algorithm for Solving Eq. (3) (max-player)

1: **Input:** Policy μ^k used in episode k, constrained policy space $\Pi_{\max}^{k+1}$, learning rate η^{k+1}, regularizer ψ, loss estimator $\widehat{\ell}^k$, number of iterations T.
2: **Initialize:** $\mu^{(1)}=\mu^k$, $\phi(\mu)=\eta^{k+1}\langle\mu,\widehat{\ell}^k\rangle+D_\psi(\mu,\mu^k)$.
3: **for** $t=1,\ldots,T$ **do**
4: Compute $g^{(t)}=\nabla\phi(\mu^{(t)})$.
5: Compute $\widehat{\mu}^{(t)}=\arg\min_{\mu\in\Pi_{\max}^{k+1}}\langle\mu,g^{(t)}\rangle$ by Algorithm 4.
6: Let $\delta=\frac{2}{1+t}$.
7: Update $\mu^{(t+1)}=(1-\delta)\mu^{(t)}+\delta\widehat{\mu}^{(t)}$.
8: **end for**
9: **Return** $\mu^{(T)}$.

In this section, we provide Algorithm 3 and Algorithm 4, which compute an approximate solution to Eq. (3).

Algorithm 4 Computing Linear Minimizer in Algorithm 3 (max-player)

1: **Input:** $\Pi_{\max}^{k+1}$, $g^{(t)}$.
2: **Initialize:** $G^{(t)}(x_h, a_h) = 0$, $\mu(a_h|x_h) = 0$, $\forall (x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$, $\forall h \in [H]$.
3: **for** $h = H, \ldots, 1$ **do**
4: **for** $x_h \in \mathcal{X}_H$ **do**
5: Compute

$$G^{(t)}(x_h, a_h) = \sum_{x_{h+1} \in C(x_h, a_h), a_{h+1} \in \mathcal{A}} \mu(a_{h+1}|x_{h+1}) \left(g^{(t)}(x_{h+1}, a_{h+1}) + G^{(t)}(x_{h+1}, a_{h+1}) \right).$$

6: Set $\mu(a_h|x_h) = \frac{1}{A(k+1)}$, $\forall a_h \in \mathcal{A}$.
7: Set $\mu(a'_h|x_h) = 1 - \frac{A-1}{A(k+1)}$, where $a'_h = \arg\min_{a \in \mathcal{A}} g^{(t)}(x_h, a) + G^{(t)}(x_h, a)$.
8: **end for**
9: **end for**
10: **Return** μ.

Computation Complexity Suppose there are K episodes. Let $w = \max_{h \in [H], (x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} |C(x_h, a_h)|$, where $C(x_h, a_h)$ is the set of immediate descendant infosets of (x_h, a_h) as defined in Section 3. Then the computation complexity of our Algorithm 2 and Algorithm 1 will be of $\mathcal{O}(wXA)$ and of $\mathcal{O}(wXA + K(XA + \text{Oracle}))$, where Oracle denotes the computation complexity of an oracle algorithm to solve our Eq. (3). If Algorithm 3 and Algorithm 4 are adopted to solve an approximate solution to Eq. (3), then Oracle will be of $\mathcal{O}(wXAT)$ where T is the number of iterations in Algorithm 3 and the total computation complexity of our Algorithm 1 will be of $\mathcal{O}(wXATK)$.

G EXPERIMENTS

In this section, we present the empirical evaluations of our Algorithm 1. Since we are not aware of any other algorithm that can also learn the (approximate) NE policy profile in IIEFGs with provable *last-iterate convergence* guarantees under bandit feedback, we compare our algorithm against previous algorithms that converge to the (approximate) NE policy profile in IIEFGs with only *average-iterate convergence* guarantees including IXOMD (Kozuno et al., 2021), BalancedOMD (Bai et al., 2022) and BalancedFTRL (Fiegel et al., 2023). Since these algorithms are only devised to obtain the average-iterate convergence for learning IIEFGs, the last-iterate convergence of these algorithms for learning IIEFGs is not theoretically guaranteed.

Environments We consider four standard IIEFG instances including Lewis Signaling, Kuhn Poker (Kuhn, 1950), Leduc Poker (Southey et al., 2012) and Liars Dice. All the implementation of these games are from the OpenSpiel library (Lanctot et al., 2019).

Implementation Details For our algorithm, to save the computation costs, instead of using our Algorithm 3 and Algorithm 4 to solve Eq. (3) in Algorithm 1, we use a lazy update of our Algorithm 1, where only the policy of the experienced trajectory of infoset action pairs $\{(x_h^k, a_h^k)\}_{h \in [H]}$ in each episode k are updated. For the remaining infoset action pairs that are not experienced by the max-player in episode k, the losses contributed by the entropy regularization (*i.e.*, the second term in our constructed entropy regularized loss estimator) of these infoset action pairs will be accumulated and will be used to update these infoset action pairs once they are experienced in some future episode, coming from the observation that the losses contributed by the entropy regularization are much smaller than the importance-weighted losses constructed using the rewards in the game (*i.e.*, the first term in our constructed entropy regularized loss estimator). In this way, the resulting computation complexity of our algorithm will only be of $\mathcal{O}(wXA + KXA)$ for running our algorithm in K episodes where $w = \max_{h \in [H], (x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} |C(x_h, a_h)|$ ($C(x_h, a_h)$ is the set of immediate descendant infosets of (x_h, a_h) as defined in Section 3). We adopt the implementation of all the baselines by Fiegel et al. (2023).[2] Besides, we consider a (logarithmic) grid search on the learning

[2] https://github.com/anon17893/IIG-tree-adaptation.

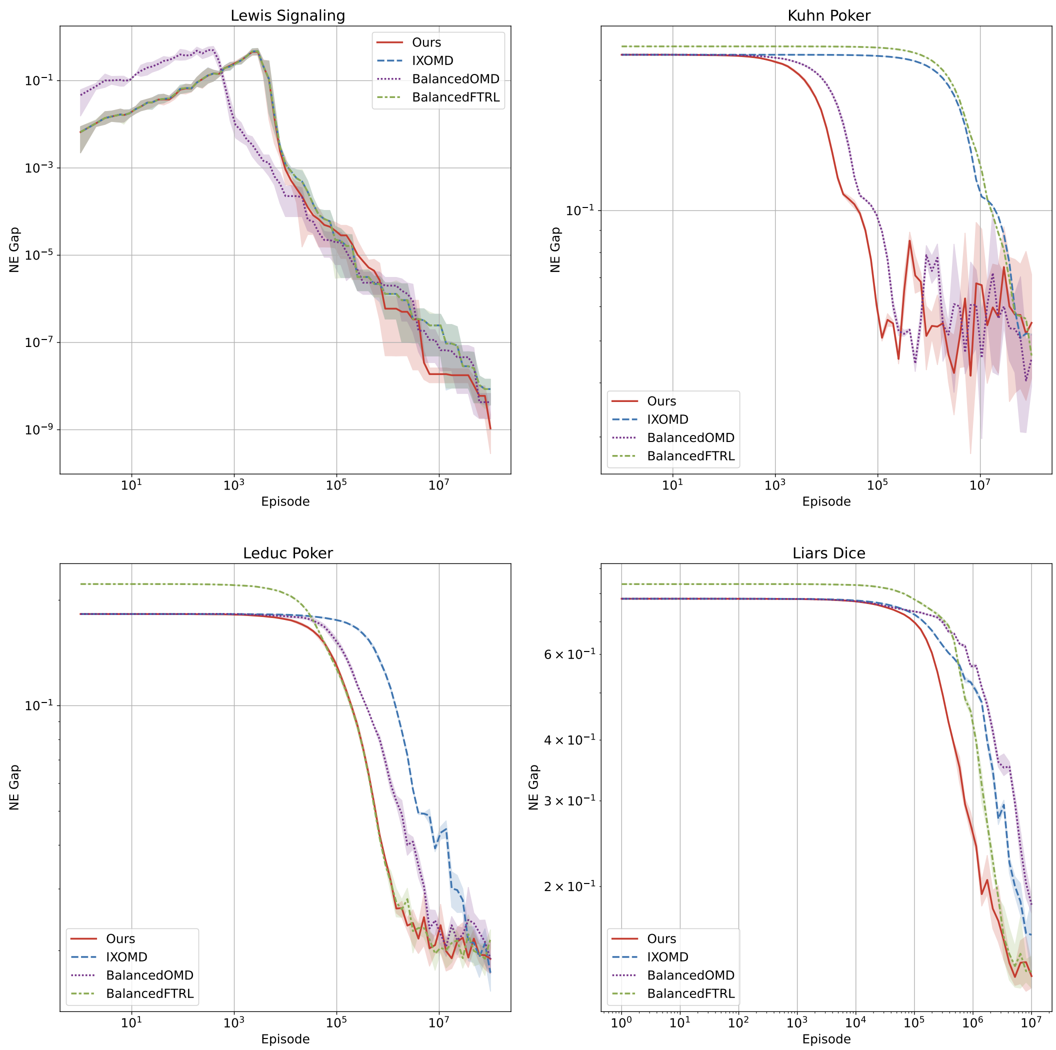

Figure 2: Experiment results of our Algorithm 1 against IXOMD (Kozuno et al., 2021), BalancedOMD (Bai et al., 2022) and BalancedFTRL (Fiegel et al., 2023). The curves show the last-iterate convergence results of the NE gap defined in Eq. (2) against the number of episodes and are averaged over 5 different seeds.

rates for all the algorithms, following Fiegel et al. (2023). All the experiments are conducted on a server with an Intel Xeon Gold CPU and 251GiB system memory. The running of all the algorithms including our algorithm costs approximately 10 hours, 12 hours, 13 hours, and 16 hours on Lewis Signaling, Kuhn Poker, Leduc Poker, and Liars Dice, respectively.

Results The experimental results are shown in Figure 2. Our algorithm obtains the best or the competitive performance across all four IIEFG instances. In particular, our algorithm converges faster than all the baseline algorithms on Kuhn Poker and Liars Dice and also converges as fast as the empirically best baseline algorithm on Lewis Signaling and Leduc Poker. Though some baseline algorithms work relatively well on some game instances, we would like to note again that these algorithms are not theoretically guaranteed to converge to the NE policy profile with the last-iterate convergence. We speculate that this might also be the reason why some baseline algorithms perform relatively well in some instances but poorly in the remaining ones. For instance, the BalancedFTRL algorithm performs well on Leduc Poker while converging very slowly on Kuhn Poker. Analogously, BalancedOMD converges relatively well on Kuhn Poker and Leduc Poker but converges the most slowly on Liars Dice.

Moreover, in general, it appears that the advantage of our algorithm becomes more pronounced in IIEFG instances with larger infoset spaces $\mathcal{X}$ (and action spaces $\mathcal{A}$) over previous algorithms. This observation aligns with the intuition that in such instances, the baseline algorithms, which solely have average-iterate convergence theoretical guarantees, face greater difficulty in achieving last-iterate convergence to the NE. This challenge may arise because these algorithms are more susceptible to getting stuck in suboptimal policy profiles, due to lack of the last-iterate convergence theoretical guarantees.