Sample-Efficient Learning of Correlated Equilibria in Extensive-Form Games

Ziang Song
Stanford University
ziangs@stanford.edu

Song Mei
UC Berkeley
songmei@berkeley.edu

Yu Bai
Salesforce Research
yu.bai@salesforce.com

Abstract

Imperfect-Information Extensive-Form Games (IIEFGs) is a prevalent model for real-world games involving imperfect information and sequential plays. The Extensive-Form Correlated Equilibrium (EFCE) has been proposed as a natural solution concept for multi-player general-sum IIEFGs. However, existing algorithms for finding an EFCE require full feedback from the game, and it remains open how to efficiently learn the EFCE in the more challenging bandit feedback setting where the game can only be learned by observations from repeated playing. This paper presents the first sample-efficient algorithm for learning the EFCE from bandit feedback. We begin by proposing K-EFCE—a generalized definition that allows players to observe and deviate from the recommended actions for K times. The K-EFCE includes the EFCE as a special case at $K = 1$, and is an increasingly stricter notion of equilibrium as K increases. We then design an uncoupled no-regret algorithm that finds an ε-approximate K-EFCE within $\widetilde{\mathcal{O}}(\max_i X_i A_i^K/\varepsilon^2)$ iterations in the full feedback setting, where X_i and A_i are the number of information sets and actions for the i-th player. Our algorithm works by minimizing a wide-range regret at each information set that takes into account all possible recommendation histories. Finally, we design a sample-based variant of our algorithm that learns an ε-approximate K-EFCE within $\widetilde{\mathcal{O}}(\max_i X_i A_i^{K+1}/\varepsilon^2)$ episodes of play in the bandit feedback setting. When specialized to $K = 1$, this gives the first sample-efficient algorithm for learning EFCE from bandit feedback.

1 Introduction

This paper is concerned with the problem of learning equilibria in Imperfect-Information Extensive-Form Games (IIEFGs) [29]. IIEFGs is a general formulation for multi-player games with both imperfect information (such as private information) and sequential play, and has been used for modeling and solving real-world games such as Poker [23, 32, 7, 8], Bridge [39], Scotland Yard [37], and so on. In a two-player zero-sum IIEFG, the standard solution concept is the celebrated notion of Nash Equilibrium (NE) [35], that is, a pair of independent policies for both players such that no player can gain by deviating. However, in multi-player general-sum IIEFGs, computing an (approximate) NE is PPAD-hard and unlikely to admit efficient algorithms [12]. A more amenable class of solution concepts is the notion of *correlated equilibria* [4], that is, a correlated policy for all players such that no player can gain by deviating from the correlated play using certain types of deviations.

The notion of Extensive-Form Correlated Equilibria (EFCE) proposed by Von Stengel and Forges [40] is a natural definition of correlated equilibria in multi-player general-sum IIEFGs. An EFCE is a correlated policy that can be thought of as a "mediator" of the game who recommends actions to each player privately and sequentially (at visited information sets), in a way that disincentivizes any player to deviate from the recommended actions. Polynomial-time algorithms for computing EFCEs have been established, by formulating as a linear program and using the ellipsoid method [24, 36, 26],

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

min-max optimization [16], or uncoupled no-regret dynamics using variants of counterfactual regret minimization [11, 20, 34, 2].

However, all the above algorithms require that the full game is known (or *full feedback*). In the more challenging *bandit feedback* setting where the game can only be learned by observations from repeated playing, it remains open how to learn EFCEs sample-efficiently. This is in contrast to other types of equilibria such as NE in two-player zero-sum IIEFGs, where sample-efficient learning algorithms under bandit feedback are known [30, 19, 28, 5]. A related question is about the structure of the EFCE definition: An EFCE only allows players to deviate *once* from the observed recommendations, upon which further recommendations are no longer revealed and the player needs to make decisions on her own instead. This may be too restrictive to model situations where players can still observe the recommendations after deviating [34]. It is of interest how we can extend the EFCE definition in a structured fashion to disincentivize such stronger deviations while still allowing efficient algorithms.

This paper makes steps towards answering both questions above, by proposing stronger and more generalized definition of EFCEs, and designing efficient learning algorithms under both full-feedback and bandit-feedback settings. We consider IIEFGs with m players, H steps, where each player i has X_i information sets and A_i actions. Our contributions can be summarized as follows.

- We propose K-EFCE, a natural generalization of EFCE, at which players have no gains when allowed to observe and deviate from the recommended actions for K times (Section 3). At $K = 1$, the K-EFCE is equivalent to the existing definition of EFCE based on trigger policies. For $K \geq 1$, the K-EFCE are increasingly stricter notions of equilibria as K increases.
- We design an algorithm K Extensive-Form Regret Minimization (K-EFR) which finds an ε-approximate K-EFCE within $\widetilde{\mathcal{O}}(\max_i X_i A_i^{K\wedge H}/\varepsilon^2)$ iterations under full feedback (Section 4). At $K = 1$, our linear in $\max_i X_i$ dependence improves over the best known $\widetilde{\mathcal{O}}(\max_i X_i^2)$ dependence for computing ε-approximate EFCE. At $K > 1$, this gives a sharp result for efficiently computing (the stricter) K-EFCE, improving over the best known $\widetilde{\mathcal{O}}(\max_i X_i^2 A_i^{K\wedge H}/\varepsilon^2)$ iteration complexity of Morrill et al. [34].
- We further design Balanced K-EFR—a sample-based variant of K-EFR—for the more challenging bandit-feedback setting (Section 5). Balanced K-EFR learns an ε-approximate K-EFCE within $\widetilde{\mathcal{O}}(\max_i X_i A_i^{K\wedge H+1}/\varepsilon^2)$ episodes of play. This is the first line of results for learning EFCE and K-EFCE from bandit feedback, and the linear in X_i dependence matches information-theoretic lower bounds. Technically, our bandit-feedback result builds on a novel stochastic wide range regret minimization algorithm SWRHEDGE, as well as sample-based estimators of counterfactual loss functions using newly designed sampling policies, which may be of independent interest.

1.1 Related work

Computing Correlated Equilibria from full feedback The notion of Extensive-Form Correlated Equilibria (EFCE) in IIEFGs is introduced in Von Stengel and Forges [40]. Huang and von Stengel [24] design the first polynomial time algorithm for computing EFCEs in multi-player IIEFGs from full feedback, using a variation of the *Ellipsoid against hope* algorithm [36, 26]. Farina et al. [16] later propose a min-max optimization formulation of EFCEs which can be solved by first-order methods.

Celli et al. [11] and its extended version [20] design the first uncoupled no-regret algorithm for computing EFCEs. Their algorithms are based on minimizing the trigger regret (first considered in Dudik and Gordon [13], Gordon et al. [22]) via counterfactual regret decomposition [45]. Morrill et al. [34] propose the stronger definition of "Behavioral Correlated Equilibria" (BCE) using general "behavioral deviations", and design the Extensive-Form Regret Minimization (EFR) algorithm to compute a BCE by using a generalized version of counterfactual regret decomposition. They also propose intermediate notions such as a "Twice Informed Partial Sequence" (TIPS) (and its K-shot generalization) as an interpolation between the EFCE and BCE. Our definition of K-EFCE offers a new interpolation between the EFCE and BCE that is different from theirs, as the deviating player in K-EFCE does not observe and does not follow recommended actions after K deviations has happened, whereas the deviator in K-shot Informed Partial Sequence resumes to following after K deviations has happened. The iteration complexity for computing an ε-approximate correlated equilibrium in both [20, 34] scales quadratically in $\max_{i\in[m]} X_i$. Our K-EFR algorithm for the full feedback setting builds upon the EFR algorithm, but specializes to the notion of K-EFCE, and achieve an improved linear in $\max_{i\in[m]} X_i$ iteration complexity.

Apart from the EFCE, there are other notions of (coarse) correlated equilibria in IIEFGs such as Normal-Form Coarse-Correlated Equilibria (NFCCE) [45, 10, 9, 18, 44], Extensive-Form Coarse-Correlated Equilibria (EFCCE) [17], and Agent-Form (Coarse-)Correlated Equilibria (AF(C)CE) [38, 40]; see [33, 34] for a detailed comparison. All above notions are either weaker than or incomparable with the EFCE, and thus results there do not imply results for computing EFCE.

Learning Equilibria from bandit feedback A line of work considers learning Nash Equilibria (NE) in two-player zero-sum IIEFGs and NFCCE in multi-player general-sum IIEFGs from bandit feedback [30, 19, 15, 21, 43, 41, 28, 5]; Note that the NFCCE is weaker than (and does not imply results for learning) EFCE. Dudík and Gordon [14] consider sample-based learning of EFCE in succinct extensive-form games; however, their algorithm relies on an approximate Markov-Chain Monte-Carlo (MCMC) sampling subroutine that does not lead to an end-to-end sample complexity guarantee. To our best knowledge, our results are the first for learning the EFCE (and K-EFCE) under bandit feedback.

2 Preliminaries

We formulate IIEFGs as partially-observable Markov games (POMGs) with tree structure and perfect recall, following [28, 5]. For a positive integer i, we denote by $[i]$ the set $\{1, 2, \cdots, i\}$. For a finite set $\mathcal{A}$, we let $\Delta(\mathcal{A})$ denote the probability simplex over $\mathcal{A}$. Let $\binom{n}{m}$ denote the binomial coefficient (i.e. number of combinations) of choosing m elements from n different elements, with the convention that $\binom{n}{m} = 0$ if $m > n$. We use $m \wedge n$ to denote $\min\{m, n\}$.

Partially observable Markov game We consider an episodic, tabular, m-player, general-sum, partially observable Markov game

$$\mathrm{POMG}(m, \mathcal{S}, \{\mathcal{X}_i\}_{i\in[m]}, \{\mathcal{A}_i\}_{i\in[m]}, H, p_0, \{p_h\}_{h\in[H]}, \{r_{i,h}\}_{i\in[m],h\in[H]}),$$

where $\mathcal{S}$ is the state space of size $|\mathcal{S}| = S$; $\mathcal{X}_i$ is the space of information sets (henceforth *infosets*) for the i^{th} player, which is a partition of $\mathcal{S}$ (i.e., $x_i \subseteq \mathcal{S}$ for all $x_i \in \mathcal{X}_i$, and $\mathcal{S} = \sqcup_{x_i\in\mathcal{X}_i} x_i$ where $\sqcup$ stands for disjoint union) with size $|\mathcal{X}_i| = X_i$, and we also use $x_i : \mathcal{S} \to \mathcal{X}_i$ to denote the i^{th} player's emission (observation) function; $\mathcal{A}_i$ is the action space for the i^{th} player with size $|\mathcal{A}_i| = A_i$, and we let $\mathcal{A} = \mathcal{A}_1 \times \cdots \mathcal{A}_m$ denote the space of joint actions $\mathbf{a} = (a_1, \dots, a_m)$; $H \in \mathbb{Z}_{\geq 1}$ is the time horizon; $p_0 \in \Delta(\mathcal{S})$ is the distribution of the initial state s_1; $p_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ are transition probabilities where $p_h(s_{h+1}|s_h, \mathbf{a}_h)$ is the probability of transiting to the next state s_{h+1} from state-action $(s_h, \mathbf{a}_h) \in \mathcal{S} \times \mathcal{A}$; and $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic[1] reward function for the i^{th} player at step h.

Tree structure and perfect recall assumption We use a POMG with tree structure and perfect recall to formulate imperfect information games[2,3], following [28, 5]. We assume that the game has a tree structure: for any state $s \in \mathcal{S}$, there is a unique step h and history $(s_1, \mathbf{a}_1, \dots, s_{h-1}, \mathbf{a}_{h-1}, s_h = s)$ to reach s. Precisely, for any policy of the players, for any realization of the game (i.e., trajectory) $(s'_k, \mathbf{a}'_k)_{k\in[H]}$, conditionally to $s'_l = s$, it almost surely holds that $l = h$ and $(s'_1, \dots, s'_h) = (s_1, \dots, s_h)$. We also assume perfect recall, which means that each player remembers its past information sets and actions. In particular, for each information set (infoset) $x_i \in \mathcal{X}_i$ for the i^{th} player, there is a unique history of past infosets and actions $(x_{i,1}, a_{i,1}, \dots, x_{i,h-1}, a_{i,h-1}, x_{i,h} = x_i)$ leading to x_i. This requires that $\mathcal{X}_i$ can be partitioned to H subsets $(\mathcal{X}_{i,h})_{h\in[H]}$ such that $x_{i,h} \in \mathcal{X}_{i,h}$ is reachable only at time step h. We define $X_{i,h} := |\mathcal{X}_{i,h}|$. Similarly, the state set $\mathcal{S}$ can be also partitioned into H subsets $(S_h)_{h\in[H]}$. As we mostly focus on the i^{th} player, we use x_h to also denote $x_{i,h}$ and use them interchangeably.

[1] Our results can be directly extended to the case of stochastic rewards.

[2] We remark that POMGs with tree structure and perfect recall is a specific subclass of POMGs; General POMGs could possess different challenges and are beyond the scope of this paper.

[3] Our definition of POMGs with tree structure and perfect recall can express any IIEFG satisfying an additional mild condition called *timeability* [25]. Further, our algorithms and guarantees can be generalized directly to any general IIEFG that is not necessarily timeable; see Appendix H.3 for discussions.

For $s \in \mathcal{S}$ and $x_i \in \mathcal{X}_i$, we write $s \in x_i$ if infoset x_i contains the state s. With an abuse of notation, for $s \in \mathcal{S}$, we let $x_i(s)$ denote the i^{th} player's infoset that s belongs to. For any $h < h'$, $x_{i,h} \in \mathcal{X}_{i,h}, x_{i,h'} \in \mathcal{X}_{i,h'}$, we write $x_{i,h} \prec x_{i,h'}$ if the information set $x_{i,h'}$ can be reached from information set $x_{i,h}$ by i^{th} player's actions; we write $(x_{i,h}, a_{i,h}) \prec x_{i,h'}$ if the infoset $x_{i,h'}$ can be reached from infoset $x_{i,h}$ by i^{th} player's action $a_{i,h}$. For any $h < h'$ and $x_{i,h} \in \mathcal{X}_{i,h}$, we let $\mathcal{C}_{h'}(x_{i,h}, a_{i,h}) := \{x \in \mathcal{X}_{i,h'} : (x_{i,h}, a_{i,h}) \prec x\}$ and $\mathcal{C}_{h'}(x_{i,h}) := \{x \in \mathcal{X}_{i,h'} : x_{i,h} \prec x\} = \cup_{a_{i,h} \in \mathcal{A}_i} \mathcal{C}_{h'}(x_{i,h}, a_{i,h})$ denote the infosets within the h'-th step that are reachable from (i.e. children of) $x_{i,h}$ or $(x_{i,h}, a_{i,h})$, respectively. For shorthand, let $\mathcal{C}(x_{i,h}, a_{i,h}) := \mathcal{C}_{h+1}(x_{i,h}, a_{i,h})$ and $\mathcal{C}(x_{i,h}) := \mathcal{C}_{h+1}(x_{i,h})$ denote the set of immediate children.

Policies We use $\pi_i = \{\pi_{i,h}(\cdot|x_{i,h})\}_{h \in [H], x_{i,h} \in \mathcal{X}_{i,h}}$ to denote a policy of the i^{th} player, where each $\pi_{i,h}(\cdot|x_{i,h}) \in \Delta(\mathcal{A}_i)$ is the action distribution at infoset $x_{i,h}$. We say π_i is a pure policy if $\pi_{i,h}(\cdot|x_{i,h})$ takes some single action deterministically for any $(h, x_{i,h})$; in this case we let $\pi_i(x_{i,h}) = \pi_{i,h}(x_{i,h})$ denote the action taken at infoset $x_{i,h}$ for shorthand. We use $\pi = \{\pi_i\}_{i \in [m]}$ to denote a product policy for all players, and let $\pi_{-i} = \{\pi_j\}_{j \in [m], j \neq i}$ denote policies of all players other than the i^{th} player. We call π a pure product policy if π_i is a pure policy for all $i \in [m]$. Let Π_i denote the set of all possible policies for the i^{th} player and $\Pi = \prod_{i \in [m]} \Pi_i$ denote the set of all possible product policies. Any probability measure $\overline{\pi}$ on Π induces a *correlated policy*, which executes as first sampling a product policy $\pi = \{\pi_i\}_{i \in [m]} \in \Pi$ from probability measure $\overline{\pi}$ and then playing the product policy π. We also use $\overline{\pi}$ to denote this policy. A correlated policy $\overline{\pi}$ can be viewed as a *mediator* of the game which samples $\pi \sim \overline{\pi}$ before the game starts, and privately *recommends* action sampled from $\pi_i(\cdot|x_i)$ to the i^{th} player when infoset $x_i \in \mathcal{X}_i$ is visited during the game.

Reaching probability With the tree structure assumption, for any state $s_h \in \mathcal{S}_h$ and actions $\mathbf{a} \in \mathcal{A}$, there exists a unique history $(s_1, \mathbf{a}_1, \ldots, s_h = s, \mathbf{a}_h = \mathbf{a})$ ending with $(s_h = s, \mathbf{a}_h = \mathbf{a})$. Given any product policy π, the probability of reaching $(s_h, \mathbf{a}_h)$ at step h can be decomposed as

$$p_h^{\pi}(s_h, \mathbf{a}) = p_{1:h}(s_h) \prod_{i \in [m]} \pi_{i,1:h}(s_h, a_{i,h}), \tag{1}$$

where we define the *sequence-form transitions* $p_{1:h}$ and *sequence-form policies* $\pi_{i,1:h}$ as

$$p_{1:h}(s_h) := \; p_0(s_1) \prod_{h'=1}^{h-1} p_{h'}(s_{h'+1}|s_{h'}, \mathbf{a}_{h'}), \tag{2}$$

$$\pi_{i,1:h}(s_h, a_{i,h}) := \; \pi_{i,1:h}(x_{i,h}, a_{i,h}) := \prod_{h'=1}^{h} \pi_{i,h'}(a_{i,h'}|x_{i,h'}), \tag{3}$$

where $(s_{h'}, \mathbf{a}_{h'})_{h' \leq h-1}$ is the unique history of states and actions that leads to s_h by the tree structure; $x_{i,h} = x_i(s_h)$ is the i^{th} player's infoset at the h-th step, and $(x_{i,h'}, a_{i,h'})_{h' \leq h-1}$ is the unique history of infosets and actions that leads to $x_{i,h}$ by perfect recall. We also define $\pi_{i,h:h'}(x_{i,h'}, a_{i,h'}) := \prod_{h''=h}^{h'} \pi_{i,h''}(a_{i,h''}|x_{i,h''})$ for any $1 \leq h \leq h' \leq H$.

Value functions and counterfactual loss functions Let $V_i^{\pi} := \mathbb{E}_{\pi}[\sum_{h=1}^{H} r_{i,h}]$ denote the value function (i.e. expected cumulative reward) for the i^{th} player under policy π. By the product form of the reaching probability in (1), the value function V_i^{π} admits a multi-linear structure over the sequence-form policies. Concretely, fixing any sequence of product policies $\{\pi^t\}_{t=1}^{T}$ where each $\pi^t = \{\pi_i^t\}_{i \in [m]}$, we have

$$V_i^{\pi^t} = \sum_{h=1}^{H} \sum_{(s_h, \mathbf{a}_h = (a_{j,h})_{j \in [m]}) \in \mathcal{S}_h \times \mathcal{A}} p_{1:h}(s_h) \prod_{j=1}^{m} \pi_{j,1:h}^t(x_j(s_h), a_{j,h}) r_{i,h}(s_h, \mathbf{a}_h).$$

For any sequence of policies $\{\pi^t\}_{t=1}^{T}$, we also define the *counterfactual loss functions* [45] $\{L_{i,h}^t(x_{i,h}, a_{i,h})\}_{i,h,x_{i,h},a_{i,h}}$ as:

$$\ell_{i,h}^t(x_{i,h}, a_{i,h}) := \sum_{\substack{s_h \in x_{i,h}, \\ \mathbf{a}_{-i,h} \in \mathcal{A}_{-i}}} p_{1:h}(s_h) \prod_{j \neq i} \pi_{j,1:h}^t(x_j(s_h), a_{j,h})[1 - r_{i,h}(s_h, \mathbf{a}_h)], \tag{4}$$

$$L_{i,h}^t(x_{i,h}, a_{i,h}) := \ell_{i,h}^t(x_{i,h}, a_{i,h}) + \sum_{h'=h+1}^{H} \sum_{\substack{x_{h'} \in \mathcal{C}_{h'}(x_{i,h}, a_{i,h}), \\ a_{h'} \in \mathcal{A}_i}} \pi_{i,(h+1):h'}^t(x_{h'}, a_{h'}) \ell_{i,h'}^t(x_{h'}, a_{h'}). \tag{5}$$

Algorithm 1 Executing modified policy $\phi \diamond \pi_i$

Input: K-EFCE strategy modification $\phi \in \Phi_i^K$ ($0 \le K \le \infty$), policy $\pi_i \in \Pi_i$ for the i^{th} player.
1: Initialize recommendation history $\mathbf{b} = \emptyset$.
2: **for** $h = 1, \dots, H$ **do**
3: Receive infoset $x_{i,h} \in \mathcal{X}_{i,h}$.
4: **if** $\mathbf{b} \in \Omega_i^{(\text{I}),K}(x_{i,h})$ **then**
5: Observe recommendation $b_h \sim \pi_{i,h}(\cdot|x_{i,h})$.
6: Take swapped action $a_h = \phi(x_{i,h}, \mathbf{b}, b_h)$.
7: Update recommendation history $\mathbf{b} \leftarrow (\mathbf{b}, b_h) \in \mathcal{A}_i^h$.
8: **else**
9: // Must have $\mathbf{b} \in \Omega_i^{(\text{II}),K}(x_{i,h})$, do not observe recommendation from π_i
10: Take action $a_h = \phi(x_{i,h}, \mathbf{b})$.

Intuitively, $L_{i,h}^t(x_{i,h}, a_{i,h})$ measures the i^{th} player's expected cumulative loss (one minus reward) conditioned on reaching $(x_{i,h}, a_{i,h})$, weighted by the (environment) transitions and all other players' policies π_{-i}^t at all time steps, and the i^{th} player's own policy π_i^t from step $h+1$ onward. We will omit the i subscript and use L_h^t to denote the above when clear from the context.

Feedback protocol We consider two standard feedback protocols for our algorithms: *full feedback*, and *bandit feedback*. In the full feedback case, the algorithm can query a product policy $\pi^t = \{\pi_i^t\}_{i\in[m]}$ in each iteration and observe the counterfactual loss functions $\{L_{i,h}^t(x_{i,h}, a_{i,h})\}_{i,h,x_{i,h},a_{i,h}}$ *exactly*[4]. In the bandit feedback case, the players can only play repeated episodes with some policies and observe the trajectory of their own infosets and rewards from the environment.

3 K Extensive-Form Correlated Equilibria

We now introduce the definition of K Extensive-Form Correlated Equilibria (K-EFCE) and establish its relationship with existing notions of correlated equilibria in IIEFGs.

3.1 Definition of K-EFCE

Intuitively, a K-EFCE is a correlated policy in which no player can gain if allowed to deviate from the observed recommended actions K times, and forced to choose her own actions without observing further recommendations afterwards. To state its definition formally, letting $\mathcal{A}_i^h := \{(b_1, \dots, b_h) | b_{h'} \in \mathcal{A}_i, \forall h' \le h\}$, we categorize all possible *recommendation histories* (henceforth *rechistories*) at each infoset $x_{i,h} \in \mathcal{X}_{i,h}$ (for the i^{th} player) into two types, based on whether the player has already deviated K times from past recommendations:

(1) A *Type-I rechistory* ($\le K-1$ deviations happened) at $x_{i,h}$ is any action sequence $b_{1:h-1} \in \mathcal{A}_i^{h-1}$ such that $\sum_{k=1}^{h-1} \mathbf{1}\{a_k \neq b_k\} \le K-1$, where $(a_1, \dots, a_{h-1})$ is the unique sequence of actions leading to $x_{i,h}$. Let $\Omega_i^{(\text{I}),K}(x_{i,h})$ denote the set of all Type-I rechistories at $x_{i,h}$.

(2) A *Type-II rechistory* (K deviations happened) at $x_{i,h}$ is any action sequence $b_{1:h'} \in \mathcal{A}_i^{h'}$ with length $h' < h$ such that $\sum_{k=1}^{h'-1} \mathbf{1}\{a_k \neq b_k\} = K-1$ and $a_{h'} \neq b_{h'}$, where $(a_1, \dots, a_{h-1})$ is the unique sequence of actions leading to $x_{i,h}$. Let $\Omega_i^{(\text{II}),K}(x_{i,h})$ denote the set of all Type-II rechistories at $x_{i,h}$.

We now define a K-EFCE strategy modification ($0 \le K \le \infty$) for the i^{th} player.

Definition 1 (K-EFCE strategy modification)**.** *A K-EFCE strategy modification ϕ (for the i^{th} player) is a mapping ϕ of the following form: At any infoset $x_{i,h} \in \mathcal{X}_{i,h}$, for any Type-I rechistory $b_{1:h-1} \in \Omega_i^{(\text{I}),K}(x_{i,h})$, ϕ swaps any recommended action b_h into $\phi(x_{i,h}, b_{1:h-1}, b_h) \in \mathcal{A}_i$; for any Type-II rechistory $b_{1:h'} \in \Omega_i^{(\text{II}),K}(x_{i,h})$, ϕ directly takes action $\phi(x_{i,h}, b_{1:h'}) \in \mathcal{A}_i$.*

[4] This is implementable (and slightly more general than) when the full game (transitions and rewards) is known.

Let Φ_i^K denote the set of all possible K-EFCE strategy modifications for any $0 \le K \le \infty$. Formally, for any $\phi \in \Phi_i^K$ and any pure policy $\pi_i \in \Pi_i$, we define the modified policy $\phi \diamond \pi_i$ as in Algorithm 1.

We parse the modified policy $\phi \diamond \pi_i$ (Algorithm 1) as follows. Upon receiving the infoset $x_{i,h}$ at each step h, the player has the rechistory $\mathbf{b}$ containing all past observed recommended actions. Then, if $\mathbf{b}$ is Type-I, i.e. at most $K-1$ deviations have happened (Line 4), then the player observes the current recommended action $b_h = \pi_{i,h}(x_{i,h})$, takes a potentially swapped action $a_h = \phi(x_{i,h}, \mathbf{b}, b_h)$ (Line 6), and appends b_h to the recommendation history (Line 7). Otherwise, $(x_{i,h}, \mathbf{b})$ is Type-II, i.e. K deviations have already happened. In this case, the player does not observe the recommended action, and instead takes an action $a_h = \phi(x_{i,h}, \mathbf{b})$, and does not update $\mathbf{b}$ (Line 10).

We now define K-EFCE as the equilibrium induced by the K-EFCE strategy modification set Φ_i^K. With slight abuse of notation, we define $\phi \diamond \overline{\pi}$ for any *correlated policy* $\overline{\pi}$ to be the policy $(\phi \diamond \pi_i) \times \pi_{-i}$ where $\pi \sim \overline{\pi}$ is the product policy sampled from $\overline{\pi}$.

Definition 2 (K-EFCE)**.** *A correlated policy $\overline{\pi}$ is an ε-approximate K Extensive-Form Correlated Equilibrium (K-EFCE) if*

$$K\text{-EFCEGap}(\overline{\pi}) := \max_{i \in [m]} \max_{\phi \in \Phi_i^K} \left(V_i^{\phi \diamond \overline{\pi}} - V_i^{\overline{\pi}} \right) \le \varepsilon.$$

We say $\overline{\pi}$ is an (exact) K-EFCE if K-EFCEGap$(\overline{\pi}) = 0$.

3.2 Properties of K-EFCE

The K-EFCE is closely related to various existing definitions of correlated equilibria in IIEFGs. We show that the special case of $K = 1$ is equivalent to the existing definition of EFCE based on trigger policies (Proposition C.1); The K-EFCE are indeed stricter equilibria as K increases (Proposition C.2); The two extreme cases $K = 0$ and $K = \infty$ are equivalent to (Normal-Form) Coarse Correlated Equilibrium and the "Behavioral Correlated Equilibria" of [34][5], respectively (Proposition C.3). Due to the space limit, the full statements and proofs are deferred to Appendix C.

4 Computing K-EFCE from full feedback

Algorithm description We first present our algorithm for computing K-EFCE in the full-feedback setting. Our algorithm K Extensive-Form Regret Minimization (K-EFR), described in Algorithm 2, is an uncoupled no-regret algorithm aiming to minimize the following *K-EFCE regret*

$$R_{i,K}^T := \max_{\phi \in \Phi_i^K} \sum_{t=1}^T \left(V_i^{\phi \diamond \pi_i^t \times \pi_{-i}^t} - V_i^{\pi^t} \right). \tag{6}$$

By standard online-to-batch conversion, achieving sublinear K-EFCE regret for every player implies that the average joint policy over all players is an approximate K-EFCE (Lemma E.1).

At a high level, our Algorithm 2 builds upon the EFR algorithm of Morrill et al. [34] to minimize the K-EFCE regret $R_{i,K}^T$, by maintaining a *regret minimizer* $\mathcal{R}_{x_{i,h}}$ (using algorithm RegAlg) at each infoset $x_{i,h} \in \mathcal{X}_i$ that is responsible for outputting the policy $\pi_i^t(\cdot|x_{i,h}) \in \Delta_{\mathcal{A}_i}$ (Line 8) which combine to give the overall policy π_i^t for the t-th iteration.

Core to our algorithm is the requirement that $\mathcal{R}_{x_{i,h}} \sim$ RegAlg should be able to minimize regrets with *time-selection functions and strategy modifications* (also known as the *wide range regret*) [31, 6]. Specifically, $\mathcal{R}_{x_{i,h}}$ needs to control the regret

$$\max_{\varphi \in \Psi^s} \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big(\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}) - \varphi \diamond \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h}, \cdot) \Big\rangle \Big) \tag{7}$$

for all possible Type-I rechistories $b_{1:h-1} \in \Omega_i^{(\mathrm{I}),K}(x_{i,h})$ simultaneously, where $\prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) =: S_{b_{1:h-1}}^t$ is the *time-selection function* (i.e. a weight function) associated with this $b_{1:h-1}$ (cf. Line 5), and $\Psi^s = \{\psi : \mathcal{A}_i \to \mathcal{A}_i\}$ is the set of all *swap modifications* from the action set $\mathcal{A}_i$ onto itself. (An analogous regret for Type-II rechistories is also controlled by $\mathcal{R}_{x_{i,h}}$.) Controlling these "local"

[5] Up to a minor difference that our ∞-EFCE only defines the equilibrium in terms of the overall game value, where the BCE additionally requries similar equilibrium properties to hold in certain subgames.

Algorithm 2 K-EFR with full feedback (i^{th} player's version)

Input: Algorithm REGALG for minimizing wide range regret; learning rates $\{\eta_{x_{i,h}}\}_{x_{i,h}\in\mathcal{X}_i}$.

1: Initialize regret minimizers $\{\mathcal{R}_{x_{i,h}}\}_{x_{i,h}\in\mathcal{X}_i}$ with REGALG and learning rate $\eta_{x_{i,h}}$.
2: **for** iteration $t = 1, \dots, T$ **do**
3:     **for** $h = 1, \dots, H$ **do**
4:         **for** $x_{i,h} \in \mathcal{X}_{i,h}$ **do**
5:             Compute $S^t_{b_{1:h-1}} = \prod_{k=1}^{h-1} \pi^t_{i,k}(b_k|x_k)$ for all $b_{1:h-1} \in \Omega_i^{(\mathrm{I}),K}(x_{i,h})$.
6:             Compute $S^t_{b_{1:h'}} = \prod_{k=1}^{h'} \pi^t_{i,k}(b_k|x_k)$ for all $b_{1:h'} \in \Omega_i^{(\mathrm{II}),K}(x_{i,h})$.
7:             $\mathcal{R}_{x_{i,h}}$.OBSERVE_TIMESELECTION$(\{S^t_{b_{1:h-1}}\}_{b_{1:h-1}\in\Omega_i^{(\mathrm{I}),K}(x_{i,h})} \cup \{S^t_{b_{1:h'}}\}_{b_{1:h'}\in\Omega_i^{(\mathrm{II}),K}(x_{i,h})})$.
8:             Set policy $\pi^t_i(\cdot|x_{i,h}) \leftarrow \mathcal{R}_{x_{i,h}}$.RECOMMEND().
9:     Observe counterfactual losses $\left\{L^t_h(x_{i,h}, a_h)\right\}_{h,x_{i,h},a_h}$ (depending on π^t_i and π^t_{-i}; cf. (5)).
10:    **for** all $x_{i,h} \in \mathcal{X}_i$ **do**
11:        $\mathcal{R}_{x_{i,h}}$.OBSERVE_LOSS$(\left\{L^t_h(x_{i,h}, a)\right\}_{a\in\mathcal{A}_i})$.

Output: Policies $\{\pi^t_i\}_{t=1}^T$.

regrets at each $x_{i,h}$ guarantees that the overall K-EFCE regret is bounded, by the K-EFCE regret decomposition (cf. Lemma E.2).

To control this wide range regret, we instantiate REGALG as WRHEDGE (Algorithm 4; cf. Appendix A.2), which is similar as the wide regret minimization algorithm in [27], with a slight modification of the initial weights suitable for our purpose (cf. (11)). The learning rate is set as

$$\eta_{x_{i,h}} = \sqrt{\binom{H}{K\wedge H} X_i A_i^{K\wedge H} \log A_i / (H^2 T)} \tag{8}$$

for all $x_{i,h} \in \mathcal{X}_i$. With this algorithm in place, at each iteration, $\mathcal{R}_{x_{i,h}}$ observes all time selection functions (Line 7), computes the policy for the current iteration (Line 8), and then observes the loss vector $L^t_{i,h}(x_{i,h}, \cdot)$ (Line 9) that is useful for updating the policy in the next iteration.

Theoretical guarantee We are now ready to present the theoretical guarantee for K-EFR.

Theorem 3 (Computing K-EFCE from full feedback)**.** *For any $0 \le K \le \infty$, $\varepsilon \in (0, H]$, let all players run Algorithm 2 together in a self-play fashion where* REGALG *is instantiated as Algorithm 4 with learning rates specified in (8). Let $\pi^t = \{\pi^t_i\}_{i\in[m]}$ denote the joint policy of all players at the t'th iteration. Then the average policy $\overline{\pi} = \mathrm{Unif}(\{\pi^t\}_{t=1}^T)$ satisfies K-$\mathrm{EFCEGap}(\overline{\pi}) \le \varepsilon$, as long as the number of iterations*

$$T \ge \mathcal{O}\Big(\binom{H}{K\wedge H}\big(\max_{i\in[m]} X_i A_i^{K\wedge H}\big)\iota/\varepsilon^2\Big),$$

where $\iota = \log(\max_{i\in[m]} A_i)$ is a log factor and $\mathcal{O}(\cdot)$ hides $\mathrm{poly}(H)$ factors.

In the special case of $K = 1$, Theorem 3 shows that K-EFR can compute an ε-approximate 1-EFCE within $\widetilde{\mathcal{O}}(\max_{i\in[m]} X_i A_i/\varepsilon^2)$ iterations. This improves over the existing $\widetilde{\mathcal{O}}(\max_{i\in[m]} X_i^2 A_i^2/\varepsilon^2)$ iteration complexity of Celli et al. [11], Farina et al. [20] by a factor of $X_i A_i$. Also, compared with the iteration complexity of the optimistic algorithm of [3] which is at least $\widetilde{\mathcal{O}}(\max_{i\in[m]} X_i^{4-\delta} A_i^{4/3}/\varepsilon^{4/3})$[6], we achieve lower X_i dependence (though worse ε dependence).

For $1 < K \le \infty$, Theorem 3 gives a sharp $\widetilde{\mathcal{O}}(\binom{H}{K\wedge H}(\max_{i\in[m]} X_i A_i^{K\wedge H})/\varepsilon^2)$ iteration complexity for computing K-EFCE. This improves over the $\widetilde{\mathcal{O}}(\binom{H}{K\wedge H}(\max_{i\in[m]} X_i^2 A_i^{K\wedge H})/\varepsilon^2)$ rate of EFR [34] instantiated to the K-EFCE problem. Also, note that although the term $A_i^{K\wedge H}$ is exponential in K (for $K \le H$), this is sensible since it is roughly the same scale as the number of possible recommendation histories, which is also the "degree of freedom" within a K-EFCE strategy modification. Apart from learning equilibria, Algorithm 2 also achieves a low K-EFCE regret when controlling the i^{th} player only and facing potentially adversarial opponents:

[6] More precisely, Anagnostides et al. [3, Corollary 4.17] proves an $\widetilde{\mathcal{O}}((X_i \max_{\pi_i\in\Pi_{\max}} \|\pi_i\|_1^2 A/\varepsilon)^{4/3})$ iteration complexity, which specializes to the above rate, as for any $\delta > 0$ a game with $\max_{\pi_i\in\Pi_{\max}} \|\pi_i\|_1 \ge X_i^{1-\delta}$ can be constructed.

Algorithm 3 Loss estimator for Type-II rechistories via Balanced Sampling (i^{th} player's version)

Input: Policy π_i^t, π_{-i}^t. Balanced exploration policies $\{\pi_i^{\star,h}\}_{h\in[H]}$.

for $K \le h' < h \le H$, $W \subseteq [h']$ with $|W| = K$ and ending in h' **do**

  Set policy $\pi_i^{t,(h,h',W)} \leftarrow (\pi_{i,k}^{\star,h})_{k\in W\cup\{h'+1,\dots,h\}} \cdot (\pi_{i,k}^t)_{k\in[h']\setminus W} \cdot \pi_{i,(h+1):H}^t$.

  Play $\pi_i^{t,(h,h',W)} \times \pi_{-i}^t$ for one episode, observe trajectory

$$(x_{i,1}^{t,(h,h',W)}, a_{i,1}^{t,(h,h',W)}, r_{i,1}^{t,(h,h',W)}, \dots, x_{i,H}^{t,(h,h',W)}, a_{i,H}^{t,(h,h',W)}, r_{i,H}^{t,(h,h',W)}).$$

  for all $(x_{i,h}, b_{1:h'}) \in \Omega_i^{(\text{II}),K}$ **do**

    Find $(x_{i,1}, a_1) \prec \cdots \prec (x_{i,h-1}, a_{h-1}) \prec x_{i,h}$.

    Set $W \leftarrow \{k \in [h'] : b_k \ne a_k\}$

    Construct loss estimator for all $a \in \mathcal{A}_i$

$$\widetilde{L}_{(x_{i,h},b_{1:h'})}^t(a) \leftarrow \frac{\mathbf{1}\left\{(x_{i,h}^{t,(h,h',W)}, a_{i,h}^{t,(h,h',W)}) = (x_{i,h}, a)\right\}}{\pi_{i,1:h}^{t,(h,h',W)}(x_{i,h}, a)} \cdot \sum_{h''=h}^{H} \left(1 - r_{i,h''}^{t,(h,h',W)}\right). \tag{9}$$

Output: Loss estimators $\left\{\widetilde{L}_{(x_{i,h},b_{1:h'})}^t(\cdot)\right\}_{(x_{i,h},b_{1:h'})\in\Omega_i^{(\text{II}),K}}$.

$R_{i,K}^T \le \widetilde{\mathcal{O}}(\sqrt{\binom{H}{K\wedge H} X_i A_i^{K\wedge H} T})$ (Corollary F.1). In particular, the $\widetilde{\mathcal{O}}(\sqrt{X_i T})$ scaling is optimal up to log factors, due to the fact that $R_{i,K}^T \ge R_{i,0}^T$ (i.e. the vanilla regret) and the known lower bound $R_{i,0}^T \ge \Omega(\sqrt{X_i T})$ in IIEFGs [42].

Proof overview Our Theorem 3 follows from a sharp analysis on the K-EFCE regret of Algorithm 2, by incorporating (i) a decomposition of the K-EFCEGap into local regrets at each infoset with tight leading coefficients (Lemma E.2), and (ii) loss-dependent upper bounds for the wide range regret of WRHEDGE (Lemma A.2), which when plugged into the aforementioned regret decomposition yields the improved dependence in $(X_i A_i^{K\wedge H})$ over the analysis of Morrill et al. [34] (Lemma F.1 & F.2), and also the $X_i A_i$ factor improvement over the results of [11, 20] in the special case of $K = 1$. The full proof can be found in Appendix F.

5 Learning K-EFCE from bandit feedback

We now present Balanced K-EFR, a sample-based variant of K-EFR that achieves a sharp sample complexity in the more challenging bandit feedback setting. Our algorithm relies on the following *balanced exploration policy* [19, 5]. Recall that $|\mathcal{C}_h(x_{i,h'}, a_{i,h'})|$ is the number of descendants of $(x_{i,h'}, a_{i,h'})$ within the h-th layer of the i^{th} player's game tree (cf. Section 2).

Definition 4 (Balanced exploration policy)**.** *For any $1 \le h \le H$, we define $\pi_i^{\star,h}$, the (i^{th} player's)* balanced exploration policy for layer h *as*

$$\pi_{i,h'}^{\star,h}(a_{h'}|x_{h'}) := |\mathcal{C}_h(x_{i,h'}, a_{i,h'})| / |\mathcal{C}_h(x_{i,h'})| \quad \textit{for all} \quad (x_{i,h'}, a_{i,h'}) \in \mathcal{X}_{i,h'} \times \mathcal{A}_i, \quad h' \le h-1,$$

and $\pi_{i,h'}^{\star,h}(a_{i,h'}|x_{i,h'}) := 1/A_i$ *for* $h' \ge h$.

Note that there are H such policies, one for each layer h. We remark that the construction of $\pi_i^{\star,h}$ requires knowledge about the descendant relationships among the i^{th} player's infosets, which is a mild requirement (e.g. can be efficiently obtained from one traversal of the i^{th} player's game tree; see Appendix H.2 for a detailed discussion about this requirement).

Algorithm description (sampling part) Our Balanced K-EFR (deferred to Algorithm 7) builds upon the full feedback version of K-EFR (Algorithm 2). The main new ingredient within Algorithm 7 is to use sample-based loss estimators obtained by two *balanced sampling* algorithms (Algorithm 3 & 6), one for each type of rechistories. Here we present the sampling algorithm for Type-II rechistories in Algorithm 3; The sampling algorithm for Type-I rechistories (Algorithm 6) is designed similarly and deferred to Appendix G.1 due to space limit. Algorithm 3 performs two main steps:

- Line 1-3 (Sampling): Construct policies $\{\pi_i^{t,(h,h',W)}\}$ that are *interlaced concatenations* of the current π_i^t and the balanced policy $\pi_i^{\star,h}$, and play one episode using each policy against π_{-i}^t.
- Line 7: Construct loss estimators $\{\widetilde{L}_{x_{i,h},b_{1:h'}}(a)\}_{x_{i,h},b_{1:h'},a}$ by (9), which for each $x_{i,h}$ and $b_{1:h'} \in \Omega_i^{(\mathrm{II}),K}(x_{i,h})$ is an unbiased estimator of counterfactual losses $\{L_h^t(x_{i,h},a)\}_{a\in\mathcal{A}_i}$ that will be used by Algorithm 7 to be fed into the regret minimization algorithm REGALG.

We remark that the sampling policies $\{\pi_i^{t,(h,h',W)}\}$ in Algorithm 3 are *interlaced concatenations* of π_i^t and $\pi_i^{\star,h}$ along time steps h, where the policy to take at each h is determined by W. These policies are generalizations of the sampling policies in the Balanced CFR algorithm of Bai et al. [5] (which can be thought of as a simple non-interlacing concatenation). They allow *time-selection aware sampling*: Each loss estimator $\widetilde{L}^t_{(x_{i,h},b_{1:h'})}(\cdot)$ achieves low variance relative to the corresponding time selection function $S^t_{b_{1:h'}}$. Further, there is an *efficient sharing* of sampling policies, as here roughly $\binom{H}{K\wedge H}X_iA_i^{K\wedge H}$ loss estimators (one for each $(x_{i,h},b_{1:h'})$) are constructed using only (a much lower number of) $H\binom{H}{K\wedge H}$ policies.

Stochastic wide-range regret minimization Algorithm 7 requires the wide-range regret minimization algorithm REGALG to additionally handle the stochastic setting, i.e. minimize the wide-range regret (e.g. (7)) when fed with our sample-based loss estimators. Here, we instantiate REGALG to be SWRHEDGE (Algorithm 5), a stochastic variant of WRHEDGE, with hyperparameters

$$\eta_{x_{i,h}} = \sqrt{\tbinom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\log(8\textstyle\sum_{i\in[m]}X_iA_i/p)/(H^3T)}, \quad \overline{L} = H. \tag{10}$$

SWRHEDGE is a non-trivial extension of WRHEDGE to the stochastic setting, as in each round it admits *multiple* sample-based loss estimators, one for each time selection function, with the same mean (cf. Line 8). This is needed since Algorithm 3 uses different sampling policies to construct the loss estimator $\widetilde{L}^t_{x_{i,h},b_{1:h'}}(\cdot)$ for each $b_{1:h'} \in \Omega_i^{(\mathrm{II}),K}(x_{i,h})$ (cf. (9)).

Theoretical guarantee We now present our main result for the bandit feedback setting.

Theorem 5 (Learning K-EFCE from bandit feedback)**.** *For any $0 \le K \le \infty$, $\varepsilon \in (0,H]$ and $p \in [0,1)$, letting all players run Algorithm 7 together in a self-play fashion for T iterations, with* REGALG *instantiated as* SWRHEDGE *(Algorithm 5) with hyperparameters in* (10)*. Let $\pi^t = \{\pi_i^t\}_{i\in[m]}$ denote the joint policy of all players at the t'th iteration. Then, with probability at least $1-p$, the correlated policy $\overline{\pi} = \mathrm{Unif}(\{\pi^t\}_{t=1}^T)$ satisfies K-*EFCEGap$(\overline{\pi}) \le \varepsilon$*, as long as $T \ge \mathcal{O}(H^3\binom{H}{K\wedge H}(\max_{i\in[m]}X_iA_i^{K\wedge H+1})\iota/\varepsilon^2)$. The total number of episodes played is*

$$3mH\tbinom{H}{K\wedge H}\cdot T = \mathcal{O}\Big(m\tbinom{H}{K\wedge H}^2\big(\max_{i\in[m]}X_iA_i^{K\wedge H+1}\big)\iota/\varepsilon^2\Big),$$

where $\iota = \log(8\sum_{i\in[m]}X_iA_i/p)$ is a log factor and $\mathcal{O}(\cdot)$ hides $\mathrm{poly}(H)$ *factors.*

To our best knowledge, Theorem 5 provides the first result for learning K-EFCE under bandit feedback. The sample complexity $\widetilde{\mathcal{O}}(\binom{H}{K\wedge H}^2\max_{i\in[m]}(X_iA_i^{K\wedge H+1})/\varepsilon^2)$ (ignoring m, H factors) has only an $\binom{H}{K\wedge H}A_i$ additional factor over the iteration complexity in the full feedback setting (Theorem 3), which is natural—The $\binom{H}{K\wedge H}$ comes from the number of episodes sampled within each iteration (Lemma G.1), and the A_i arises from estimating loss vectors from bandit feedback. In particular, the special case of $K=1$ provides the first result for learning EFCEs from bandit feedback, with sample complexity $\widetilde{\mathcal{O}}(\max_{i\in[m]}X_iA_i^2/\varepsilon^2)$. We remark that the linear in X_i dependence at all $K \ge 0$ is optimal, as the sample complexity lower bound for the $K=0$ case (learning NFCCEs) is already $\Omega(\max_{i\in[m]}X_iA_i/\varepsilon^2)$ [5][7]. Also, the policies $\{\pi_i^t\}_{t=1}^T$ maintained in Algorithm 7 also achieves sublinear K-EFCE regret. However, strictly speaking, this is not a regret bound of our algorithm, as the sampling policies $\pi_i^{t,(h,h',W)}$ actually used are not π_i^t.

Proof overview The proof of Theorem 5 builds on the analysis in the full-feedback case, and further relies on several new techniques in order to achieve the sharp linear in $\max_{i\in[m]}X_i$ sample

[7] The sample complexity lower bound in [5] is stated for learning Nash Equilibria in two-player zero-sum IIEFGs, but can be directly extended to learning NFCCEs in multi-player general-sum IIEFGs.

complexity: (1) A regret bound for the SWRHEDGE algorithm under the same-mean condition (Lemma A.3), which may be of independent interest; (2) Crucial use of the *balancing property* of $\pi_i^{\star,h}$ (Lemma B.4) to control the variance of the loss estimators $\widetilde{L}^t_{(x_{i,h},b_{1:h'})}(\cdot)$, which in turn produces sharp bounds on the regret terms and additional concentration terms (Lemma G.4-G.9). The full proof can be found in Appendix G.3.

6 Conclusion

This paper proposes K-EFCE, a generalized definition of Extensive-Form Correlated Equilibria in Imperfect-Information Games, and designs sharp algorithms for computing K-EFCE under full feedback and learning a K-EFCE under bandit feedback. Our algorithms perform wide-range regret minimization over each infoset to minimize the overall K-EFCE regret, and introduce new efficient sampling policies to handle bandit feedback. We believe our work opens up many future directions, such as accelerated techniques for computing K-EFCE from full feedback, learning other notions of equilibria from bandit feedback, as well as empirical investigations of our algorithms.

Acknowledgment

The authors thank Brian Hu Zhang for discussions regarding the relationship between the ∞-EFCE and the BCE. S.M. is supported by NSF grant DMS-2210827.

References

[1] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[2] Ioannis Anagnostides, Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster no-regret learning dynamics for extensive-form correlated equilibrium. 2021.

[3] Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Andrea Celli, and Tuomas Sandholm. Faster no-regret learning dynamics for extensive-form correlated and coarse correlated equilibria. *arXiv preprint arXiv:2202.05446*, 2022.

[4] Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

[5] Yu Bai, Chi Jin, Song Mei, and Tiancheng Yu. Near-optimal learning of extensive-form games with imperfect information. *arXiv preprint arXiv:2202.01752*, 2022.

[6] Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

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

[8] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365 (6456):885–890, 2019.

[9] Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting cfr+ and alternating updates. *Journal of Artificial Intelligence Research*, 64:429–443, 2019.

[10] Andrea Celli, Alberto Marchesi, Tommaso Bianchi, and Nicola Gatti. Learning to correlate in multi-player general-sum sequential games. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Andrea Celli, Alberto Marchesi, Gabriele Farina, and Nicola Gatti. No-regret learning dynamics for extensive-form correlated equilibrium. *arXiv preprint arXiv:2004.00603*, 2020.

[12] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[13] Miroslav Dudik and Geoffrey Gordon. A sampling-based approach to computing equilibria in succinct extensive-form games. *arXiv preprint arXiv:1205.2649*, 2012.

[14] Miroslav Dudík and Geoffrey J Gordon. A sampling-based approach to computing equilibria in succinct extensive-form games. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 151–160, 2009.

[15] Gabriele Farina and Tuomas Sandholm. Model-free online learning in unknown sequential decision making problems and games. *arXiv preprint arXiv:2103.04539*, 2021.

[16] Gabriele Farina, Chun Kai Ling, Fei Fang, and Tuomas Sandholm. Correlation in extensive-form games: Saddle-point formulation and benchmarks. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Gabriele Farina, Tommaso Bianchi, and Tuomas Sandholm. Coarse correlation in extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1934–1941, 2020.

[18] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. *arXiv preprint arXiv:2007.14358*, 2020.

[19] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Stochastic regret minimization in extensive-form games. In *International Conference on Machine Learning*, pages 3018–3028. PMLR, 2020.

[20] Gabriele Farina, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium. *arXiv preprint arXiv:2104.01520*, 2021.

[21] Gabriele Farina, Robin Schmucker, and Tuomas Sandholm. Bandit linear optimization for sequential decision making and extensive-form games. *arXiv preprint arXiv:2103.04546*, 2021.

[22] Geoffrey J Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, pages 360–367, 2008.

[23] Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *International conference on machine learning*, pages 805–813. PMLR, 2015.

[24] Wan Huang and Bernhard von Stengel. Computing an extensive-form correlated equilibrium in polynomial time. In *International Workshop on Internet and Network Economics*, pages 506–513. Springer, 2008.

[25] Sune K Jakobsen, Troels B Sørensen, and Vincent Conitzer. Timeability of extensive-form games. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 191–199, 2016.

[26] Albert Xin Jiang and Kevin Leyton-Brown. Polynomial-time computation of exact correlated equilibrium in compact games. *Games and Economic Behavior*, 91:347–359, 2015.

[27] Subhash Khot and Ashok Kumar Ponnuswami. Minimizing wide range regret with time selection functions. In *COLT*, pages 81–86, 2008.

[28] Tadashi Kozuno, Pierre Ménard, Rémi Munos, and Michal Valko. Model-free learning for two-player zero-sum partially observable markov games with perfect recall. *arXiv preprint arXiv:2106.06279*, 2021.

[29] Harold W Kuhn. Extensive games and the problem of information. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 193–216. Princeton University Press, 1953.

[30] Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael H Bowling. Monte carlo sampling for regret minimization in extensive games. In *NIPS*, pages 1078–1086, 2009.

[31] Ehud Lehrer. A wide range no-regret theorem. *Games and Economic Behavior*, 42(1):101–115, 2003.

[32] 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.

[33] Dustin Morrill, Ryan D'Orazio, Reca Sarfati, Marc Lanctot, James R Wright, Amy Greenwald, and Michael Bowling. Hindsight and sequential rationality of correlated play. *arXiv preprint arXiv:2012.05874*, 2020.

[34] Dustin Morrill, Ryan D'Orazio, Marc Lanctot, James R Wright, Michael Bowling, and Amy R Greenwald. Efficient deviation types and learning for hindsight rationality in extensive-form games. In *International Conference on Machine Learning*, pages 7818–7828. PMLR, 2021.

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

[36] Christos H Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)*, 55(3):1–29, 2008.

[37] Martin Schmid, Matej Moravcik, Neil Burch, Rudolf Kadlec, Josh Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, Zach Holland, et al. Player of games. *arXiv preprint arXiv:2112.03178*, 2021.

[38] Reinhard Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4, 1975.

[39] Yuandong Tian, Qucheng Gong, and Tina Jiang. Joint policy search for multi-agent collaboration with imperfect information. *arXiv preprint arXiv:2008.06495*, 2020.

[40] Bernhard Von Stengel and Françoise Forges. Extensive-form correlated equilibrium: Definition and computational complexity. *Mathematics of Operations Research*, 33(4):1002–1022, 2008.

[41] Brian Hu Zhang and Tuomas Sandholm. Finding and certifying (near-) optimal strategies in black-box extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5779–5788, 2021.

[42] Yichi Zhou, Tongzheng Ren, Jialian Li, Dong Yan, and Jun Zhu. Lazy-cfr: fast and near optimal regret minimization for extensive games with imperfect information. *arXiv preprint arXiv:1810.04433*, 2018.

[43] Yichi Zhou, Jialian Li, and Jun Zhu. Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information. In *International Conference on Learning Representations*, 2019.

[44] Yichi Zhou, Tongzheng Ren, Jialian Li, Dong Yan, and Jun Zhu. Lazy-cfr: fast and near-optimal regret minimization for extensive games with imperfect information. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJx4p3NYDB.

[45] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20:1729–1736, 2007.

Appendix

Table of Contents

A Technical tools

A.1 Technical lemmas

The following Freedman's inequality can be found in [1, Lemma 9].

Lemma A.1 (Freedman's inequality)**.** *Suppose random variables $\{X_t\}_{t=1}^T$ is a martingale difference sequence, i.e. $X_t \in \mathcal{F}_t$ where $\{\mathcal{F}_t\}_{t\geq 1}$ is a filtration, and $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Suppose $X_t \leq R$ almost surely for some (non-random) $R > 0$. Then for any $\lambda \in (0, 1/R]$, we have with probability at least $1 - \delta$ that*

$$\sum_{t=1}^{T} X_t \leq \lambda \cdot \sum_{t=1}^{T} \mathbb{E}\big[X_t^2|\mathcal{F}_{t-1}\big] + \frac{\log(1/\delta)}{\lambda}.$$

A.2 Wide-range regret minimization

Time selection functions For each element b in a finite set $\mathcal{B}$, we let $\{S_b^t\}_{t\geq 1} \subseteq [0, 1]$ represent the time selection functions (viewing it as a function of t). We let $\mathcal{B}$ be the non-intersection union of $\mathcal{B}^s$ and $\mathcal{B}^e$, where $\mathcal{B}^s$ is called the set of swap time selection indexes and $\mathcal{B}^e$ is called the set of external time selection indexes.

Strategy modifications We denote the set of swap modification function set $\Psi^s = \{\psi : [A] \to [A]\}$ and external modification function set $\Psi^e = \{\psi : [A] \to [A] : \exists a \in [A], \text{s.t. } \psi(b) = a, \forall b \in [A]\}$. Given a modification $\psi : [A] \to [A]$ and a strategy $p^t \in \Delta([A])$, we denote $\psi \diamond p^t \in \Delta([A])$ to be the modified strategy with $(\psi \diamond p^t)(a) = \sum_{b\in\psi^{-1}(a)} p^t(b)$. Given a modification $\psi : [A] \to [A]$, we define $M_\psi \in \{0, 1\}^{A\times A}$ to be its associated matrix, with its $(b, \psi(b))$'th element $M_\psi(b, \psi(b)) = 1$ for every $b \in [A]$, and otherwise equal to 0.

Interaction protocal and learning goals We consider the following interaction protocol: at each iteration t, the learner receives the time selection set $\{S_b^t\}_{b\in\mathcal{B}} \subseteq [0, 1]$, outputs a vector $p^t \in \Delta([A])$, and receives the loss $\ell^t \in [0, \infty]^A$. The goal of the learner is to control the regret $\sum_{t=1}^T S_b^t(\langle p^t, \ell^t\rangle - \langle \psi \diamond p^t, \ell^t\rangle)$ for each pair $(b, \psi) \in \mathcal{I} := (\mathcal{B}^s \times \Psi^s) \cup (\mathcal{B}^e \times \Psi^e)$.

Specializing the results in [27], we design the WRHEDGE algorithm (Algorithm 4) achieving our goal. The regret bound is presented in the lemma below, whose proof is based on [27], with a slight modification on the initial weights and with a refined analysis.

Algorithm 4 Wide-Range Regret Minimization with Hedge (WRHEDGE)

Input: Learning rate $\eta > 0$. Swap index set $\mathcal{B}^s$ and external index set $\mathcal{B}^e$; Swap strategy modification set Ψ^s and external strategy modification set Ψ^e.

1: Let $\mathcal{I} := (\mathcal{B}^s \times \Psi^s) \cup (\mathcal{B}^e \times \Psi^e)$. Initialize $S_b^0 \leftarrow 0$ for all $b \in \mathcal{B}^s \cup \mathcal{B}^e$, and

$$q^0(b, \psi) \leftarrow \frac{|\Psi^s|\mathbf{1}\{b \in \mathcal{B}^e\} + |\Psi^e|\mathbf{1}\{b \in \mathcal{B}^s\}}{\sum_{(b',\psi')\in\mathcal{I}} [|\Psi^s|\mathbf{1}\{b' \in \mathcal{B}^e\} + |\Psi^e|\mathbf{1}\{b' \in \mathcal{B}^s\}]} \tag{11}$$

for all $(b, \psi) \in \mathcal{I}$.

2: **for** iteration $t = 1, \dots, T$ **do**

3: (OBSERVE_TIMESELECTION) Receive time selection functions $\{S_b^t\}_{b\in\mathcal{B}^s\cup\mathcal{B}^e}$.

4: Update distribution over $(b, \psi) \in \mathcal{I}$:

$$q^t(b, \psi) \propto q^{t-1}(b, \psi) \exp\Big\{\eta \exp(-\eta\|\ell^{t-1}\|_\infty) S_b^{t-1}\langle p^{t-1}, \ell^{t-1}\rangle - \eta S_b^{t-1}\langle \psi \diamond p^{t-1}, \ell^{t-1}\rangle\Big\}.$$

5: Set $p^t \in \Delta([A])$ as a solution to the equation $p^{t\top} = p^{t\top}\left(\frac{\sum_{(b,\psi)\in\mathcal{I}} S_b^t q^t(b,\psi) M_\psi}{\sum_{(b,\psi)\in\mathcal{I}} S_b^t q^t(b,\psi)}\right)$.

6: **if** RECOMMEND is called **then**

7: Output the vector p^t.

8: (OBSERVE_LOSS) Receive loss vector $\ell^t \in \mathbb{R}^A$.

Lemma A.2 (Wide-range regret bound of WRHEDGE). *Let $\{\ell^t(a)\}_{a\in[A],t\in[T]}$ and $\{S_b^t\}_{b\in\mathcal{B}^s\cup\mathcal{B}^e,t\in[T]}$ be arbitrary arrays of loss functions and time selection functions. Assume that $S_b^t \in [0,1]$ and $\ell^t \in [0,\infty]^A$ for any $b \in \mathcal{B}^s \cup \mathcal{B}^e$, $a \in [A]$ and $t \in [T]$. Let p^t be given as in Algorithm 4 with learning rate $\eta \in (0,\infty)$. Then we have*

$$\sup_{\psi\in\Psi^s}\sum_{t=1}^T S_b^t\Big(\left\langle p^t,\ell^t\right\rangle - \left\langle \psi\diamond p^t,\ell^t\right\rangle\Big) \le \sum_{t=1}^T \eta\|\ell^t\|_\infty S_b^t\left\langle p^t,\ell^t\right\rangle + \frac{\log[(|\mathcal{B}^s|+|\mathcal{B}^e|)|\Psi^s|]}{\eta},\quad \forall b\in\mathcal{B}^s,$$
$$\sup_{\psi\in\Psi^e}\sum_{t=1}^T S_b^t\Big(\left\langle p^t,\ell^t\right\rangle - \left\langle \psi\diamond p^t,\ell^t\right\rangle\Big) \le \sum_{t=1}^T \eta\|\ell^t\|_\infty S_b^t\left\langle p^t,\ell^t\right\rangle + \frac{\log[(|\mathcal{B}^s|+|\mathcal{B}^e|)|\Psi^e|]}{\eta},\quad \forall b\in\mathcal{B}^e.$$

Proof. For $(b,\psi)\in\mathcal{I}$, we define the cumulative loss w.r.t. $\{S_b^t\}_{t\ge1}$ till time t as

$$L_b^t := \sum_{t'=1}^t S_b^{t'}\left\langle p^{t'},\ell^{t'}\right\rangle,$$

and define the cumulative loss w.r.t. $(\{S_b^t\}_{t\ge1},\psi)$ till time t as

$$L^t(b,\psi) := \sum_{t'=1}^t S_b^{t'}\left\langle \psi\diamond p^{t'},\ell^{t'}\right\rangle.$$

We further define the weight of $(\{S_b^t\}_{t\ge1},\psi)$ at the end of time t as

$$w^t(b,\psi) := (|\Psi^s|\mathbf{1}\{b\in\mathcal{B}^e\}+|\Psi^e|\mathbf{1}\{b\in\mathcal{B}^s\})\exp\left\{\eta\sum_{t'=1}^t \exp(-\eta\|\ell^{t'}\|_\infty)S_b^{t'}\left\langle p^{t'},\ell^{t'}\right\rangle - \eta L^t(b,\psi)\right\},$$

and hence $w^0(b,\psi)$ is given by $|\Psi^s|\mathbf{1}\{b\in\mathcal{B}^e\}+|\Psi^e|\mathbf{1}\{b\in\mathcal{B}^s\}$. We further let $W^t := \sum_{(b,\psi)\in\mathcal{I}} w^t(b,\psi)$. Then the quantity $q^t(b,\psi)$ in Algorithm 4 is simply equal to $w^{t-1}(b,\psi)/W^{t-1}$.

We next show that $W^t \le W^{t-1}$ for all $t\ge1$. In fact, by $\exp(-\eta x)\le 1-(1-\exp(-\eta\|\ell^t\|_\infty))x/\|\ell^t\|_\infty$ and $\exp(\eta x)\le 1+(\exp(\eta\|\ell^t\|_\infty)-1)x/\|\ell^t\|_\infty$ for any $\eta\in(0,\infty)$ and $x\in[0,\|\ell^t\|_\infty]$, we have

$$\begin{aligned}
W^t &= \sum_{(b,\psi)\in\mathcal{I}} w^t(b,\psi) = \sum_{(b,\psi)\in\mathcal{I}} w^{t-1}(b,\psi)\exp\left\{\eta S_b^t\left\langle \exp(-\eta\|\ell^t\|_\infty)p^t - \psi\diamond p^t,\ell^t\right\rangle\right\}\\
&\le \sum_{(b,\psi)\in\mathcal{I}} w^{t-1}(b,\psi)\left(1-\frac{(1-\exp(-\eta\|\ell^t\|_\infty))S_b^t}{\|\ell^t\|_\infty}\left\langle\psi\diamond p^t,\ell^t\right\rangle\right)\cdot\left(1+\frac{(1-\exp(-\eta\|\ell^t\|_\infty))S_b^t}{\|\ell^t\|_\infty}\left\langle p^t,\ell^t\right\rangle\right)\\
&\overset{(i)}{\le} W^{t-1} - \frac{1-\exp(-\eta\|\ell^t\|_\infty)}{\|\ell^t\|_\infty}W^{t-1}\sum_{(b,\psi)\in\mathcal{I}} q^t(b,\psi)S_b^t\left\langle\psi\diamond p^t,\ell^t\right\rangle\\
&\qquad + \frac{1-\exp(-\eta\|\ell^t\|_\infty)}{\|\ell^t\|_\infty}W^{t-1}\sum_{(b,\psi)\in\mathcal{I}} q^t(b,\psi)S_b^t\left\langle p^t,\ell^t\right\rangle\\
&\overset{(ii)}{=} W^{t-1}.
\end{aligned}$$

Here, (i) follows from $q^t(b,\psi)=w^{t-1}(b,\psi)/W^{t-1}$ and $\ell^t\in[0,1]^A$; (ii) uses the fact that p^t solves

$$p^{t\top} = p^{t\top}\left(\frac{\sum_{(b,\psi)\in\mathcal{I}} S_b^t q^t(b,\psi)M_\psi}{\sum_{(b,\psi)\in\mathcal{I}} S_b^t q^t(b,\psi)}\right)$$

in line 5 in the algorithm, which gives

$$\sum_{(b,\psi)\in\mathcal{I}} q^t(b,\psi)S_b^t p^t = \sum_{(b,\psi)\in\mathcal{I}} q^t(b,\psi)S_b^t(\psi\diamond p^t).$$

This proves that $W^t\le W^{t-1}$ for all $t\ge1$.

Therefore, W^t is non-increasing in t, and thus for all $(b, \psi) \in \mathcal{I}$,

$$
\begin{aligned}
&(|\Psi^s| \mathbf{1}\{b \in \mathcal{B}^e\} + |\Psi^e| \mathbf{1}\{b \in \mathcal{B}^s\}) \exp\left\{ \eta \sum_{t'=1}^{t} \exp(-\eta \|\ell^{t'}\|_\infty) S_b^{t'} \left\langle p^{t'}, \ell^{t'} \right\rangle - \eta L^t(b, \psi) \right\} \\
&= w^t(b, \psi) \le \sum_{(b', \psi') \in \mathcal{I}} w^0(b', \psi') = |\Psi^s||\Psi^e|(|\mathcal{B}^e| + |\mathcal{B}^s|),
\end{aligned}
$$

which gives

$$
\begin{aligned}
\sum_{t'=1}^{t} \exp(-\eta \|\ell^{t'}\|_\infty) S_b^{t'} \left\langle p^{t'}, \ell^{t'} \right\rangle - L^t(b, \psi) &\le \frac{\log[|\Psi^s|(|\mathcal{B}^e| + |\mathcal{B}^s|)]}{\eta}, \quad \forall b \in \mathcal{B}^s, \\
\sum_{t'=1}^{t} \exp(-\eta \|\ell^{t'}\|_\infty) S_b^{t'} \left\langle p^{t'}, \ell^{t'} \right\rangle - L^t(b, \psi) &\le \frac{\log[|\Psi^e|(|\mathcal{B}^e| + |\mathcal{B}^s|)]}{\eta}, \quad \forall b \in \mathcal{B}^e.
\end{aligned}
$$

Note that we have $1 \le \exp(-\eta \|\ell^t\|_\infty) + \eta \|\ell^t\|_\infty$. So we can get that, for any $b \in \mathcal{B}^s$,

$$
\begin{aligned}
L^t(b) - L^t(b, \psi) &= \sum_{t=1}^{T} S_b^t \left\langle p^t, \ell^t \right\rangle - L^t(b, \psi) \\
&\le \sum_{t=1}^{T} \exp(-\eta \|\ell^t\|_\infty) S_b^t \left\langle p^{t'}, \ell^{t'} \right\rangle - L^t(b, \psi) + \sum_{t=1}^{T} \eta \|\ell^t\|_\infty S_b^t \left\langle p^t, \ell^t \right\rangle \\
&\le \sum_{t=1}^{T} \eta \|\ell^t\|_\infty S_b^t \left\langle p^t, \ell^t \right\rangle + \frac{\log[|\Psi^s|(|\mathcal{B}^e| + |\mathcal{B}^s|)]}{\eta}.
\end{aligned}
$$

Note that the left side is exactly $\sum_{t=1}^{T} S_b^t \Big(\left\langle p^t, \ell^t \right\rangle - \left\langle \psi \diamond p^t, \ell^t \right\rangle \Big)$. Consequently,

$$
\sum_{t=1}^{T} S_b^t \Big(\left\langle p^t, \ell^t \right\rangle - \left\langle \psi \diamond p^t, \ell^t \right\rangle \Big) \le \sum_{t=1}^{T} \eta \|\ell^t\|_\infty S_b^t \left\langle p^t, \ell^t \right\rangle + \frac{\log[|\Psi^s|(|\mathcal{B}^e| + |\mathcal{B}^s|)]}{\eta}, \quad \forall b \in \mathcal{B}^s.
$$

We have similar results for $b \in \mathcal{B}^e$. Taking supremum over all $\psi \in \Psi^e$ or $\psi \in \Psi^s$ proves Lemma A.2. □

A.3 Stochastic wide-range regret minimization

In this section, we consider a stochastic variant of wide-range regret minimization. More specifically, we consider the following interaction protocol: at each iteration t, the learner receives the time selection set $\{S_b^t\}_{b \in \mathcal{B}} \subseteq [0, 1]$, outputs a vector $p^t \in \Delta([A])$, and receives the unbiased stochastic loss $\widetilde{\ell}_b^t$ for each $b \in \mathcal{B}^e \cup \mathcal{B}^s$, with $\mathbb{E}[\widetilde{\ell}_b^t | \mathcal{F}_{t-1}] = \ell^t$ (the expected loss is independent of b). Here, the σ-field $\mathcal{F}_{t-1}$ is generated by all the random variables before the t-th round. The goal of the learner is still to control the regret $\sum_{t=1}^{T} S_b^t(\langle p^t, \widetilde{\ell}_b^t \rangle - \langle \psi \diamond p^t, \widetilde{\ell}_b^t \rangle)$ for each pair $(b, \psi) \in \mathcal{I} := (\mathcal{B}^s \times \Psi^s) \cup (\mathcal{B}^e \times \Psi^e)$.

Our algorithm Stochastic Wide-Range Regret Minimization with Hedge (SWRHEDGE) is given in Algorithm 5. The regret bound for SWRHEDGE is given by the lemma below.

Lemma A.3 (Wide-range regret bound for SWRHEDGE)**.** *Let $\{\ell^t\}_{t \in [T]}$, $\{M_b^t\}_{b \in \mathcal{B}^s \cup \mathcal{B}^e, t \in [T]}$, and $\{w_b\}_{b \in \mathcal{B}^s \cup \mathcal{B}^e}$ be arbitrary arrays of loss functions, rescaled time selection functions, and weighting functions. Let $0 < \overline{L} < \infty$ be a parameter (that will serve as an upper bound of all $w_b M_b^t \|\widetilde{\ell}_b^t\|_\infty$). Assume that (i) $M_b^t \ge 0$, $w_b > 0$, and $\widetilde{\ell}_b^t \in [0, \overline{L}/(w_b M_b^t)]^A$ for any $b \in \mathcal{B}^s \cup \mathcal{B}^e$, $a \in [A]$ and $t \in [T]$; (ii) $\mathbb{E}\Big[\widetilde{\ell}_b^t | \mathcal{F}_{t-1}\Big] = \ell^t$ for all $b \in \mathcal{B}^s \cup \mathcal{B}^e$. Let p^t be given as in Algorithm 5 with learning rate $\eta \in (0, \infty)$ and time selection function $\{S_b^t\}_{b \in \mathcal{B}^s \cup \mathcal{B}^e} = \{w_b M_b^t\}_{b \in \mathcal{B}^s \cup \mathcal{B}^e}$. Then with probability at least $1 - p$, we have*

$$
\begin{aligned}
\sup_{\psi \in \Psi^s} \sum_{t=1}^{T} M_b^t \Big(\left\langle p^t, \widetilde{\ell}_b^t \right\rangle - \left\langle \psi \diamond p^t, \widetilde{\ell}_b^t \right\rangle \Big) &\le \sum_{t=1}^{T} \eta \overline{L} M_b^t \left\langle p^t, \widetilde{\ell}_b^t \right\rangle + \frac{\log[(|\mathcal{B}^s| + |\mathcal{B}^e|)|\Psi^s|/p]}{\eta w_b}, \quad \forall b \in \mathcal{B}^s, \\
\sup_{\psi \in \Psi^e} \sum_{t=1}^{T} M_b^t \Big(\left\langle p^t, \widetilde{\ell}_b^t \right\rangle - \left\langle \psi \diamond p^t, \widetilde{\ell}_b^t \right\rangle \Big) &\le \sum_{t=1}^{T} \eta \overline{L} M_b^t \left\langle p^t, \widetilde{\ell}_b^t \right\rangle + \frac{\log[(|\mathcal{B}^s| + |\mathcal{B}^e|)|\Psi^e|/p]}{\eta w_b}, \quad \forall b \in \mathcal{B}^e.
\end{aligned}
$$

Algorithm 5 Stochastic Wide Range Regret Minimization with Hedge (SWRHEDGE)

Input: Learning rate $\eta > 0$ and parameter $\overline{L}$. Swap index set $\mathcal{B}^s$ and external index set $\mathcal{B}^e$; Swap strategy modification set Ψ^s and external strategy modification set Ψ^e.

1: Let $\mathcal{I} := (\mathcal{B}^s \times \Psi^s) \cup (\mathcal{B}^e \times \Psi^e)$. Initialize $S_b^0 \leftarrow 0$ for all $b \in \mathcal{B}^s \cup \mathcal{B}^e$, and

$$q^0(b,\psi) \leftarrow \frac{|\Psi^s| \mathbf{1}\{b \in \mathcal{B}^e\} + |\Psi^e| \mathbf{1}\{b \in \mathcal{B}^s\}}{\sum_{(b',\psi') \in \mathcal{I}} [|\Psi^s| \mathbf{1}\{b' \in \mathcal{B}^e\} + |\Psi^e| \mathbf{1}\{b' \in \mathcal{B}^s\}]}$$

for all $(b,\psi) \in \mathcal{I}$.

2: **for** iteration $t = 1, \ldots, T$ **do**

3: (OBSERVE_TIMESELECTION) Receive time selection functions $\{S_b^t\}_{b \in \mathcal{B}^s \cup \mathcal{B}^e}$.

4: Update distribution over $(b,\psi) \in \mathcal{I}$:

$$q^t((b,\psi)) \propto q^{t-1}((b,\psi)) \exp\left\{\eta \exp(-\eta \overline{L}) S_b^{t-1} \langle p^{t-1}, \widetilde{\ell}_b^{t-1} \rangle - \eta S_b^{t-1} \langle \psi \diamond p^{t-1}, \widetilde{\ell}_b^{t-1} \rangle \right\}.$$

5: Set $p^t \in \Delta([A])$ as a solution to the equation ${p^t}^\top = {p^t}^\top \left(\frac{\sum_{(b,\psi) \in \mathcal{I}} S_b^t q^t(b,\psi) M_\psi}{\sum_{(b,\psi) \in \mathcal{I}} S_b^t q^t(b,\psi)} \right)$.

6: **if** RECOMMEND is called **then**

7: Output the vector p^t.

8: (OBSERVE_LOSS) Receive loss vectors $\{\widetilde{\ell}_b^t\}_{b \in \mathcal{B}^s \cup \mathcal{B}^e}$ (where $\mathbb{E}\left(\widetilde{\ell}_b^t | \mathcal{F}_{t-1}\right) = \ell^t$ doesn't depend on b).

Proof. For $(b,\psi) \in \mathcal{I}$, we define the cumulative loss w.r.t. $\{S_b^t\}_{t \geq 1}$ till time t as

$$L^t(b) := \sum_{t'=1}^{t} S_b^{t'} \left\langle p^{t'}, \widetilde{\ell}_b^{t'} \right\rangle,$$

and define the cumulative loss w.r.t. $(\{S_b^t\}_{t \geq 1}, \psi)$ till time t as

$$L^t(b,\psi) := \sum_{t'=1}^{t} S_b^{t'} \left\langle \psi \diamond p^{t'}, \widetilde{\ell}_b^{t'} \right\rangle.$$

We further define the weight of $(\{S_b^t\}_{t \geq 1}, \psi)$ at the end of time t as

$$w^t(b,\psi) := (|\Psi^s| \mathbf{1}\{b \in \mathcal{B}^e\} + |\Psi^e| \mathbf{1}\{b \in \mathcal{B}^s\}) \exp\left\{\eta \exp(-\eta \overline{L}) L^t(b) - \eta L^t(b,\psi)\right\},$$

and hence $w^0(b,\psi)$ is given by $|\Psi^s| \mathbf{1}\{b \in \mathcal{B}^e\} + |\Psi^e| \mathbf{1}\{b \in \mathcal{B}^s\}$. We further let $W^t := \sum_{(b,\psi) \in \mathcal{I}} w^t(b,\psi)$. Then the quantity $q^t(b,\psi)$ in Algorithm 5 is simply equal to $w^{t-1}(b,\psi)/W^{t-1}$.

We next show that $\mathbb{E}[W^t | \mathcal{F}_{t-1}] \leq W^{t-1}$ for all $t \geq 1$. In fact, by $\exp(-\eta x) \leq 1 - (1 - \exp(-\eta \overline{L}))x/\overline{L}$ and $\exp(\eta x) \leq 1 + (\exp(\eta \overline{L}) - 1)x/\overline{L}$ for any $\eta \in (0,\infty]$ and $x \in [0, \overline{L}]$, we have

$$\begin{aligned}
W^t &= \sum_{(b,\psi) \in \mathcal{I}} w^t(b,\psi) = \sum_{(b,\psi) \in \mathcal{I}} w^{t-1}(b,\psi) \exp\left\{\eta S_b^t \left\langle \exp(-\eta \overline{L}) p^t - \psi \diamond p^t, \widetilde{\ell}_b^t \right\rangle\right\} \\
&\leq \sum_{(b,\psi) \in \mathcal{I}} w^{t-1}(b,\psi) \left(1 - \frac{(1 - \exp(-\eta \overline{L})) S_b^t}{\overline{L}} \left\langle \psi \diamond p^t, \widetilde{\ell}_b^t \right\rangle\right) \cdot \left(1 + \frac{(1 - \exp(-\eta \overline{L})) S_b^t}{\overline{L}} \left\langle p^t, \widetilde{\ell}_b^t \right\rangle\right) \\
&\overset{(i)}{\leq} W^{t-1} - \frac{1 - \exp(-\eta \overline{L})}{\overline{L}} W^{t-1} \sum_{(b,\psi) \in \mathcal{I}} q^t(b,\psi) S_b^t \left\langle \psi \diamond p^t, \widetilde{\ell}_b^t \right\rangle \\
&\quad + \frac{1 - \exp(-\eta \overline{L})}{\overline{L}} W^{t-1} \sum_{(b,\psi) \in \mathcal{I}} q^t(b,\psi) S_b^t \left\langle p^t, \widetilde{\ell}_b^t \right\rangle.
\end{aligned}$$

Here, (i) follows from $q^t(b,\psi) = w^{t-1}(b,\psi)/W^{t-1}$ and $S_b^t \|\widetilde{\ell}_b^t\|_\infty \le \overline{L}$ for any b. Using the above inequality, $\mathbb{E}\Big(\widetilde{\ell}_b^t | \mathcal{F}_{t-1}\Big) = \ell^t$ for any b, and the fact that p^t solves

$$p^{t\top} = p^{t\top} \left(\frac{\sum_{(b,\psi)\in\mathcal{I}} S_b^t q^t(b,\psi) M_\psi}{\sum_{(b,\psi)\in\mathcal{I}} S_b^t q^t(b,\psi)} \right)$$

in line 5 in the algorithm, which gives

$$\sum_{(b,\psi)\in\mathcal{I}} q^t(b,\psi) S_b^t p^t = \sum_{(b,\psi)\in\mathcal{I}} q^t(b,\psi) S_b^t (\psi \diamond p^t),$$

we have $\mathbb{E}[W^t|\mathcal{F}_{t-1}] \le W^{t-1}$. Taking expectation and using the tower property of conditional expectation yields that

$$\mathbb{E}\big[W^t\big] \le W^0.$$

Therefore, by Markov inequality, we have with probability at least $1-p$ that

$$W^t \le W^0/p.$$

On this event, we have for all $(b,\psi)\in\mathcal{I}$ that

$$\begin{aligned}&(|\Psi^s|\mathbf{1}\{b\in\mathcal{B}^e\} + |\Psi^e|\mathbf{1}\{b\in\mathcal{B}^s\}) \exp\big\{\eta\exp(-\eta\overline{L})L^t(b) - \eta L^t(b,\psi)\big\} \\ &= w^t(b,\psi) \le W^t \le W^0/p \le \sum_{(b',\psi')\in\mathcal{I}} w^0(b',\psi')/p = |\Psi^s||\Psi^e|(|\mathcal{B}^e| + |\mathcal{B}^s|)/p.\end{aligned}$$

As a result,

$$\begin{aligned}\exp\big\{\eta\exp(-\eta\overline{L})L^t(b) - \eta L^t(b,\psi)\big\} &\le |\Psi^s|(|\mathcal{B}^e| + |\mathcal{B}^s|)/p, \quad \forall b\in\mathcal{B}^s,\\ \exp\big\{\eta\exp(-\eta\overline{L})L^t(b) - \eta L^t(b,\psi)\big\} &\le |\Psi^e|(|\mathcal{B}^e| + |\mathcal{B}^s|)/p, \quad \forall b\in\mathcal{B}^e.\end{aligned}$$

Note that we have $1 \le \exp(-\eta\overline{L}) + \eta\overline{L}$. So we can get that, for any $b\in\mathcal{B}^s$,

$$\begin{aligned}L^t(b) - L^t(b,\psi) &\le \exp(-\eta\overline{L})L^t(b) - L^t(b,\psi) + \eta\overline{L}L^t(b)\\ &\le \frac{\log[|\Psi^s|(|\mathcal{B}^e| + |\mathcal{B}^s|)/p]}{\eta} + \eta\overline{L}L^t(b).\end{aligned}$$

Note that the left side is exactly $\sum_{t=1}^T S_b^t\Big(\Big\langle p^t, \widetilde{\ell}_b^t\Big\rangle - \Big\langle \psi\diamond p^t, \widetilde{\ell}_b^t\Big\rangle\Big)$. Consequently,

$$\sum_{t=1}^T S_b^t\Big(\Big\langle p^t, \widetilde{\ell}_b^t\Big\rangle - \Big\langle \psi\diamond p^t, \widetilde{\ell}_b^t\Big\rangle\Big) \le \sum_{t=1}^T \eta\overline{L}S_b^t\Big\langle p^t, \widetilde{\ell}_b^t\Big\rangle + \frac{\log[|\Psi^s|(|\mathcal{B}^e| + |\mathcal{B}^s|)/p]}{\eta}, \quad \forall b\in\mathcal{B}^s.$$

Because $S_b^t = M_b^t w_b$, dividing by w_b gives that

$$\sum_{t=1}^T M_b^t\Big(\Big\langle p^t, \widetilde{\ell}_b^t\Big\rangle - \Big\langle \psi\diamond p^t, \widetilde{\ell}_b^t\Big\rangle\Big) \le \sum_{t=1}^T \eta\overline{L}M_b^t\Big\langle p^t, \widetilde{\ell}_b^t\Big\rangle + \frac{\log[|\Psi^s|(|\mathcal{B}^e| + |\mathcal{B}^s|)/p]}{\eta w_b}, \quad \forall b\in\mathcal{B}^s.$$

We have similar results for $b\in\mathcal{B}^e$. Taking supreme over all $\psi\in\Psi^e$ or $\psi\in\Psi^s$ proves Lemma A.3. ☐

B Properties of the game

B.1 Basic properties

Given the sequence-form transitions $p_{1:h}$ as in Eq. (2) and the sequence-form policies of the opponents $\{\pi_{j,1:h}\}_{j\neq i}$ as in Eq. (3), we define the marginal reaching probability $p_{1:h}^{\pi_{-i}}(s_h)$ and $p_{1:h}^{\pi_{-i}}(x_{i,h})$ as follows:

$$p_{1:h}^{\pi_{-i}}(s_h) = p_{1:h}(s_h) \prod_{j\in[m], j\neq i} \pi_{j,1:h}(s_h, a_{j,h}), \tag{12}$$

$$p_{1:h}^{\pi_{-i}}(x_{i,h}) = \sum_{s_h\in x_{i,h}} p_{1:h}^{\pi_{-i}}(s_h). \tag{13}$$

The following three results give properties of the marginal reaching probability $p_{1:h}^{\pi_{-i}}(x_{i,h})$, and the counterfactural loss functions $L_{i,h}^t$.

Lemma B.1 (Properties of $p_{1:h}^{\pi_{-i}}(x_h)$)**.** *The following holds for any $\pi_{-i} = \{\pi_j\}_{j\neq i} \in \otimes_{j\neq i}\Pi_j$:*

(a) For any policy $\pi_i \in \Pi_i$, we have

$$\sum_{(x_h,a_h)\in\mathcal{X}_{i,h}\times\mathcal{A}_i} \pi_{i,1:h}(x_h,a_h)p_{1:h}^{\pi_{-i}}(x_h) = 1.$$

(b) $0 \le p_{1:h}^{\pi_{-i}}(x_h) \le 1$ for all $h \in [H], x_h \in \mathcal{X}_{i,h}$.

Proof. For (a), notice that

$$\pi_{i,1:h}(x_h,a_h)p_{1:h}^{\pi_{-i}}(x_h) = \sum_{s_h\in x_h} p_{1:h}(s_h)\cdot\pi_{i,1:h}(x_h,a_h)\cdot\prod_{j\neq i}\pi_{j,1:h-1}(x_{j,h}(s_{h-1}),a_{j,h-1})$$
$$= \sum_{s_h\in x_h}\mathbb{P}^{\pi_i,\pi_{-i}}(\text{visit }(s_h,a_h)) = \mathbb{P}^{\pi_i,\pi_{-i}}(\text{visit }(x_h,a_h)).$$

Summing over all $(x_h,a_h)\in\mathcal{X}_{i,h}\times\mathcal{A}_i$, the right hand side sums to one, thereby showing (a).

For (b), fix any $x_h \in \mathcal{X}_{i,h}$. Clearly $p_{1:h}^{\pi_{-i}}(x_h) \ge 0$. Choose any $a_h \in \mathcal{A}_i$, and choose policy $\pi_i^{x_h,a_h} \in \Pi_i$ such that $\pi_{i,1:h}^{x_h,a_h}(x_h,a_h) = 1$ (such $\pi_i^{x_h,a_h}$ exists, for example, by deterministically taking all actions prescribed in infoset x_h at all ancestors of x_h). For this $\pi_i^{x_h,a_h}$, using (a), we have

$$p_{1:h}^{\pi_{-i}}(x_h) = \pi_{i,1:h}^{x_h,a_h}(x_h,a_h)\cdot p_{1:h}^{\pi_{-i}}(x_h) \le \sum_{(x_h',a_h')\in\mathcal{X}_{i,h}\times\mathcal{A}_i}\pi_{i,1:h}^{x_h,a_h}(x_h',a_h')\cdot p_{1:h}^{\pi_{-i}}(x_h') = 1.$$

This shows part (b). □

Corollary B.1. *For any policy $\pi_i \in \Pi_i$ and $h \in [H]$, we have*

$$\sum_{(x_h,a_h)\in\mathcal{X}_{i,h}\times\mathcal{A}_i}\pi_{i,1:h}(x_h,a_h)\ell_{i,h}^t(x_h,a_h) \le 1.$$

Proof. Notice by definition

$$\ell_h^t(x_h,a_h) = \sum_{s_h\in x_h,(a_{j,h})_{j\neq i}\in\otimes_{j\neq i}\mathcal{A}_j} p_{1:h}(s_h)\prod_{j\neq i}\pi_{j,1:h}^t(x_{j,h}(s_h),a_{j,h})(1-r_h(s_h,\mathbf{a}_h)) \le p_{1:h}^{\pi_{-i}}(x_h),$$

and the result is implied by Lemma B.1 (b). □

Lemma B.2. *For any $h \in [H]$, the counterfactual loss function $L_{i,h}^t$ defined in (5) satisfies the bound*

(a) For any policy $\pi_i \in \Pi_i$, we have

$$\sum_{(x_h,a_h)\in\mathcal{X}_{i,h}\times\mathcal{A}_i}\pi_{i,1:h}(x_h,a_h)L_{i,h}^t(x_h,a_h) \le H-h+1.$$

(b) For any (h,x_h,a_h), we have

$$0 \le L_{i,h}^t(x_h,a_h) \le p_{1:h}^{\pi_{-i}^t}(x_h)\cdot(H-h+1).$$

Proof. Part (a) follows from the fact that

$$\sum_{(x_h,a_h)\in\mathcal{X}_{i,h}\times\mathcal{A}_i}\pi_{i,1:h}(x_h,a_h)L_{i,h}^t(x_h,a_h) = \mathbb{E}_{\pi_i,\pi_{-i}^t}\left[\sum_{h'=h}^H r_{h'}\right] \le H-h+1,$$

where the first equality follows from the definition of the loss functions ℓ_h and L_h in (4), (5).

For part (b), the nonnegativity follows clearly by definition. For the upper bound, take any policy $\pi_i^{x_h,a_h} \in \Pi_i$ such that $\pi_{i,1:h}^{x_h,a_h}(x_h,a_h) = 1$. We then have

$$L_{i,h}^t(x_h,a_h) = \pi_{i,1:h}^{x_h,a_h}(x_h,a_h)L_{i,h}^t(x_h,a_h) = \mathbb{E}_{\pi_i^{x_h,a_h},\pi_{-i}^t}\left[\mathbf{1}\{\text{visit }x_h,a_h\}\cdot\sum_{h'=h}^H r_{h'}\right]$$

$$= \mathbb{P}_{\pi_i^{x_h,a_h}, \pi_{-i}^t}(\text{visit } x_h, a_h) \cdot \mathbb{E}_{\pi_i^{x_h,a_h}, \pi_{-i}^t}\left[\sum_{h'=h}^{H} r_{h'} \middle| \text{visit } x_h, a_h \right]$$
$$\leq \pi_{i,1:h}^{x_h,a_h}(x_h, a_h) p_{1:h}^{\pi_{-i}^t}(x_h) \cdot (H - h + 1) = p_{1:h}^{\pi_{-i}^t}(x_h) \cdot (H - h + 1).$$

This proves the lemma. □

B.2 Balanced exploration policy

Here we collect properties of the balanced exploration policy $\pi_i^{\star,h}$ (cf. Definition 4). Most results below have appeared in [5, Appendix C.2] in the two-player zero-sum setting. Here we present them again in our setting of multi-player general-sum IIEFGs.

We begin by providing an interpretation of the balanced exploration policy $\pi_{i,1:h}^{\star,h}$: its inverse $1/\pi_{i,1:h}^{\star,h}$ can be viewed as the (product) of a "transition probability" over the game tree for the i'th player.

For any $1 \leq h \leq H$ and $1 \leq k \leq h-1$, define $p_{i,k}^{\star,h}(x_{k+1}|x_k, a_k) = |\mathcal{C}_h(x_{k+1})|/|\mathcal{C}_h(x_k, a_k)|$ (we use the convention that $|\mathcal{C}_h(x_h)| = 1$). By this definition, $p_{i,k}^{\star,h}(\cdot|x_k, a_k)$ is a probability distribution over $\mathcal{C}_h(x_k, a_k)$ and can be interpreted as a balanced transition probability from (x_k, a_k) to x_{k+1}. The sequence-form of this balanced transition probability takes the form

$$p_{i,1:h}^{\star,h}(x_h) = \frac{|\mathcal{C}_h(x_1)|}{X_{i,h}} \prod_{k=1}^{h-1} p_{i,k}^{\star,h}(x_{k+1}|x_k, a_k) = \frac{|\mathcal{C}_h(x_1)|}{X_{i,h}} \prod_{k=1}^{h-1} \frac{|\mathcal{C}_h(x_{k+1})|}{|\mathcal{C}_h(x_k, a_k)|}. \tag{14}$$

Lemma B.3. *For any $(x_h, a_h) \in \mathcal{X}_{i,h} \times \mathcal{A}_i$, the sequence form of the transition $p_{i,1:h}^{\star,h}(x_h)$ and the sequence form of balanced exploration policy $\pi_{i,1:h}^{\star,h}(x_h, a_h)$ are related by*

$$p_{i,1:h}^{\star,h}(x_h) = \frac{1}{X_{i,h} A_i \cdot \pi_{i,1:h}^{\star,h}(x_h, a_h)}. \tag{15}$$

Furthermore, for any i-th player's policy $\pi_i \in \Pi_i$ and any $h \in [H]$, we have

$$\sum_{(x_h,a_h) \in \mathcal{X}_{i,h} \times \mathcal{A}_i} \pi_{i,1:h}(x_h, a_h) p_{i,1:h}^{\star,h}(x_h) = 1. \tag{16}$$

Proof. By the definition of the balanced transition probability as in Eq. (14) and the balanced exploration policy as in Definition 4, we have

$$\frac{1}{X_{i,h} A_i \cdot \pi_{i,1:h}^{\star,h}(x_h, a_h)} = \frac{1}{X_{i,h} A_i} \prod_{k=1}^{h-1} \frac{|\mathcal{C}_h(x_k)|}{|\mathcal{C}_h(x_k, a_k)|} \times A_i = \frac{|\mathcal{C}_h(x_1)|}{X_{i,h}} \prod_{k=1}^{h-1} \frac{|\mathcal{C}_h(x_{k+1})|}{|\mathcal{C}_h(x_k, a_k)|} = p_{i,1:h}^{\star,h}(x_h),$$

where the second equality used the property that $|\mathcal{C}_h(x_h)| = 1$. This proves Eq. (15). The proof of Eq. (16) is similar to the proof of Lemma B.1 (a). □

Lemma B.4 (Balancing property of $\pi_i^{\star,h}$). *For any i^{th} player's policy $\pi_i \in \Pi_i$ and any $h \in [H]$, we have*

$$\sum_{(x_h,a_h) \in \mathcal{X}_{i,h} \times \mathcal{A}_i} \frac{\pi_{i,1:h}(x_h, a_h)}{\pi_{i,1:h}^{\star,h}(x_h, a_h)} = X_{i,h} A_i.$$

Proof. Lemma B.4 follows as a direct consequence of Eq. (15) and (16) in Lemma B.3. □

Lemma B.4 states that $\pi_i^{\star,h}$ is a good exploration policy in the sense that the distribution ratio between it and *any* $\pi_i \in \Pi_i$ has bounded L_1 norm. Further, the bound $X_{i,h} A_i$ is non-trivial—For example, if we replace $\pi_{i,1:h}^{\star,h}$ with the uniform policy $\pi_{i,1:h}^{\text{unif}}(x_h, a_h) = 1/A_i^h$, the left-hand side can be as large as $X_{i,h} A_i^h$ in the worst case.

C Relationship between K-EFCE and exsiting equilibria

Equivalence between 1-EFCE **and trigger definition of EFCE** At the special case $K = 1$, our (exact) 1-EFCE is equivalent to the existing definition of EFCE based on *trigger policies* [22, 11], which defines an ε-approximate EFCE as any correlated policy $\overline{\pi}$ such that the following trigger gap is at most ε:

$$\mathrm{TriggerGap}(\overline{\pi}) := \max_{i \in [m]} \max_{(x_i, a) \in \mathcal{X}_i \times \mathcal{A}_i} \max_{\widehat{\pi}_i \in \Pi_i} \left(\mathbb{E}_{\pi \sim \overline{\pi}} V_i^{\mathsf{trig}(\pi_i, \widehat{\pi}_i, (x_i, a)) \times \pi_{-i}} - \mathbb{E}_{\pi \sim \overline{\pi}} V_i^{\pi} \right) \le \varepsilon. \quad (17)$$

Here the trigger policy $\mathsf{trig}(\pi_i, \widehat{\pi}_i, (x_i, a)) \in \Pi_i$ (with triggering sequence (x_i, a)) is the unique policy that plays $\pi_i \in \Pi_i$, unless infoset x_i is visited and action a is recommended, in which case the sequence (x_i, a) is "triggered" and the player plays $\widehat{\pi}_i \in \Pi_i$ thereafter.

Proposition C.1 (Equivalence of 1-EFCE and trigger definition)**.** *For any correlated policy $\overline{\pi}$, we have*

$$\mathrm{TriggerGap}(\overline{\pi}) \le 1\text{-}\mathrm{EFCEGap}(\overline{\pi}) \le (\max_{i \in [m]} X_i A_i) \cdot \mathrm{TriggerGap}(\overline{\pi}),$$

In particular, $1\text{-}\mathrm{EFCEGap}(\overline{\pi}) = 0$ *if and only if* $\mathrm{TriggerGap}(\overline{\pi}) = 0$.

The proof can be found in Appendix C.1. Proposition C.1 has two main implications: (1) An exact EFCE defined by the trigger gap is equivalent to an exact 1-EFCE (cf. Definition 2). Therefore the two definitions yields the same set of exact equilibria. (2) For $\varepsilon > 0$, $1\text{-}\mathrm{EFCEGap}(\overline{\pi}) \le \varepsilon$ implies $\mathrm{TriggerGap}(\overline{\pi}) \le \varepsilon$, but the converse only holds with an extra $\max_{i \in [m]} X_i A_i$ factor, and thus 1-EFCEGap is a stricter metric for approximate equilibria than TriggerGap. This distinction is inherent instead of a proof artifact: Our 1-EFCE strategy modification (Algorithm 1) is able to *implement multiple trigger policies simultaneously*, as long as their triggering sequences are not ancestors or descendants of each other.

Containment relationship We next show that K-EFCE are indeed stricter equilibria as K increases, i.e. any (approximate) $(K + 1)$-EFCE is also an (approximate) K-EFCE, but not the converse. This justifies the necessity of considering K-EFCE for all values of $K \ge 1$ and shows that they are strict strengthenings of the 1-EFCE. Note that as we consider games with a finite horizon H, we have $K\text{-}\mathrm{EFCEGap} = H\text{-}\mathrm{EFCEGap}$ for all $K \ge H$ (including $K = \infty$). The proof of Proposition C.2 can be found in Appendix C.2.

Proposition C.2 (Containment relationship)**.** *For any correlated policy $\overline{\pi}$, we have*

$$\begin{aligned}&0\text{-}\mathrm{EFCEGap}(\overline{\pi}) \le 1\text{-}\mathrm{EFCEGap}(\overline{\pi}) \le \cdots \le K\text{-}\mathrm{EFCEGap}(\overline{\pi}) \le (K+1)\text{-}\mathrm{EFCEGap}(\overline{\pi}) \\ &\le \cdots \le \infty\text{-}\mathrm{EFCEGap}(\overline{\pi}).\end{aligned}$$

In other words, K-EFCE *are stricter equilibria as K increases: Any ε-approximate* $(K + 1)$-EFCE *is also an ε-approximate* K-EFCE *for any $\varepsilon \ge 0$ and $K \ge 0$.*

Moreover, the converse bounds do not hold, even if multiplicative factors are allowed: For any $0 \le K < \infty$, there exists a game with $H = K + 1$ and a correlated policy $\overline{\pi}$ for which

$$K\text{-}\mathrm{EFCEGap}(\overline{\pi}) = 0 \quad \text{but} \quad (K+1)\text{-}\mathrm{EFCEGap}(\overline{\pi}) \ge 1/3 > 0.$$

Relationship with other correlated equilibria The two endpoints $K = 0$ and $K = \infty$ of K-EFCE are closely related to other existing definitions of correlated equilibria in IIEFGs. Concretely, 0-EFCE is equivalent to Normal-Form Coarse Correlated Equilibria (NFCCE), whereas ∞-EFCE is equivalent to using the "Behavioral Correlated Equilibria" considered in [34], which is strictly weaker than Normal-Form Correlated Equilibria (NFCE) that is more computationally challenging to learn [17, 11].

In order to introduce the definition of NFCE, we reload the definition of a correlated policy to be a probability measure on all *pure product policies* instead of general product policies. We let Π_i^{pure} denote the set of all possible pure policies for player i. Note that this does not affect our definition of K-EFCE introduced in Section 3.

We first present the definitions of Normal-Form Correlated Equilibria (NFCE) and Normal-Form Coarse Correlated Equilibria (NFCCE) (from e.g. [17]). For consistency with our K-EFCE definition, we define both equilibria through defining their set of strategy modifications.

Definition C.1 (NFCE strategy modification)**.** *A* NFCE *strategy modification ϕ (for the i^{th} player) is a mapping $\phi(\cdot, \cdot) : \mathcal{X}_i \times \Pi_i^{\rm pure} \to \mathcal{A}_i$. Let $\Phi_i^{\rm NFCE}$ denote the set of all possible* NFCE *strategy modification for the i^{th} player. For any $\phi \in \Phi_i^{\rm NFCE}$, and any pure policy $\pi_i \in \Pi_i^{\rm pure}$, we define the modified policy $\phi \diamond \pi_i$ as following: at infoset $x_{i,h}$, the modified policy $\phi \diamond \pi_i$ takes action $\phi(x_{i,h}, \pi_i)$.*

Definition C.2 (NFCCE strategy modification)**.** *A* NFCCE *strategy modification ϕ (for the i^{th} player) is a mapping $\phi(\cdot) : \mathcal{X}_i \to \mathcal{A}_i$. Let $\Phi_i^{\rm NFCCE}$ denote the set of all possible* NFCCE *strategy modification for the i^{th} player. For any $\phi \in \Phi_i^{\rm NFCCE}$, and any pure policy $\pi_i \in \Pi_i^{\rm pure}$, we define the modified policy $\phi \diamond \pi_i$ as following: at infoset $x_{i,h}$, the modified policy $\phi \diamond \pi_i$ take action $\phi(x_{i,h})$.*

At a high level, NFCE has the "strongest" form of strategy modifications, which can observe the entire pure policy π_i (i.e. full set of recommendations on every infoset). NFCCE has the "weakest" form of strategy modifications, which cannot observe any recommendation at all (so that each $\phi \in \Phi_i^{\rm NFCCE}$ is equivalent to a pure policy in $\Pi_i^{\rm pure}$).

Definition C.3 (NFCE and NFCCE)**.** *An ε-approximate* $\{$NFCE, NFCCE$\}$ *of a POMG is a correlated policy $\overline{\pi}$ such that*

$$\{\rm NFCE, NFCCE\}Gap(\overline{\pi}) := \max_{i\in[m]} \max_{\phi\in\Phi_i^{\{\rm NFCE, NFCCE\}}} \left(\mathbb{E}_{\pi\sim\overline{\pi}} V_i^{(\phi\diamond\pi_i)\times\pi_{-i}} - \mathbb{E}_{\pi\sim\overline{\pi}} V_i^{\pi}\right) \le \varepsilon.$$

We say $\overline{\pi}$ is an (exact) $\{\rm NFCE, NFCCE\}$ *if the above holds with $\varepsilon = 0$.*

Proposition C.3 (Relationship between K-EFCE and NFCE, NFCCE)**.** *For any correlated policy $\overline{\pi}$, we have*

(a) ∞-EFCEGap$(\overline{\pi}) \le$ NFCEGap$(\overline{\pi})$*, i.e.* NFCE *is stricter than* ∞-EFCE (NFCEGap$(\overline{\pi}) \le \varepsilon$ *implies* ∞-EFCEGap$(\overline{\pi}) \le \varepsilon$*).*

Further, the converse bound does not hold even if multiplicative factors are allowed: there exists a game with $H = 2$ and a correlated policy $\overline{\pi}$ for which

$$\infty\text{-EFCEGap}(\overline{\pi}) = 0 \quad \text{but} \quad \text{NFCEGap}(\overline{\pi}) > 1/20.$$

(b) 0-EFCEGap$(\overline{\pi}) =$ NFCCEGap$(\overline{\pi})$*, i.e.* 0-EFCE *is equivalent to* NFCCE.

The proof can be found in Section C.3.

Equivalence between ∞-EFCE and BCE deviations Next, we give an (informal) argument of the equivalence between "behavioral deviations" considered in [34] and our ∞-EFCE strategy modifications.

A "behavioral deviation" ϕ for one player states that at each infoset, the player can choose from three options: (i) follow the recommendation action, (ii) choose a action without ever seeing the recommendation action, or (iii) choose an action after seeing the recommendation action. Further, the choice of these three options as well as the action to deviate to may depend on the infoset as well as the recommendation history. This is exactly equivalent to the ∞-EFCE strategy modification defined in Definition 1.

We remark though, despite the equivalence between the strategy modifications of ∞-EFCE and BCE, the resulting equilibria defined as the BCE in Morrill et al. [34] is slightly stricter than the ∞-EFCE—The definition of Morrill et al. [34] requires a BCE π to satisfy that π_i does not gain in game value from all the above deviation functions in not only the full game, but also in certain subgames induced by π_{-i}; by contrast, our ∞-EFCE only requires such a property in the full game.

C.1 Proof of Proposition C.1

Proof. It suffices to consider all trigger policy $\mathsf{trig}(\pi_i, \widehat{\pi}_i, (x_i, a))$ where $\widehat{\pi}_i$ is a pure policy (at each infoset $x_{h'}$, $\widehat{\pi}_i$ chooses action $\widehat{\pi}_i(x_{h'})$ deterministically.). We prove the two claims separately.

Step 1. We first show that TriggerGap$(\overline{\pi}) \le$ 1-EFCEGap$(\overline{\pi})$ for any correlated policy $\overline{\pi}$. We first claim that, for any trigger policy $\mathsf{trig}(\pi_i, \widehat{\pi}_i, (x_{i,h}^\star, a^\star))$ where $\pi_i, \widehat{\pi}_i \in \Pi_i$, $x_{i,h}^\star \in \mathcal{X}_{i,h}$, and $a^\star \in \mathcal{A}_i$, there exists an 1-EFCE strategy modification $\phi^\star \in \Phi_i^1$ such that, for any opponent's policy $\pi_{-i} \in \Pi_{-i}$, we have

$$V_i^{\mathsf{trig}(\pi_i, \widehat{\pi}_i, (x_{i,h}^\star, a^\star))\times\pi_{-i}} = V_i^{(\phi^\star\diamond\pi_i)\times\pi_{-i}}.$$

Given this claim, for any correlated policy $\overline{\pi}$, we have as desired

$$\begin{aligned}\text{TriggerGap}(\overline{\pi}) &= \max_{i\in[m]} \max_{(x_i,a)\in\mathcal{X}_i\times\mathcal{A}_i} \max_{\widehat{\pi}_i\in\Pi_i} \left(\mathbb{E}_{\pi\sim\overline{\pi}} V_i^{\mathsf{trig}(\pi_i,\widehat{\pi}_i,(x_i,a))\times\pi_{-i}} - \mathbb{E}_{\pi\sim\overline{\pi}} V_i^{\pi}\right) \\ &\le \max_{i\in[m]} \max_{\phi\in\Phi_i^1} \left(\mathbb{E}_{\pi\sim\overline{\pi}} V_i^{(\phi\diamond\pi_i)\times\pi_{-i}} - \mathbb{E}_{\pi\sim\overline{\pi}} V_i^{\pi}\right) = \text{1-EFCEGap}(\overline{\pi}).\end{aligned}$$

To prove such a claim, we can choose the 1-EFCE strategy modification $\phi^\star$ to be the following: (1) At any Type-I rechistory with infoset $x = x_{i,h}^\star$, $\phi^\star$ swaps $a_\star$ to $\widehat{\pi}_i(x_{i,h}^\star)$ and swaps a to a (i.e., keep it unchanged) for any $a \neq a_\star$; (2) At any Type-I rechistory with infoset x such that $x \neq x_{i,h}^\star$ and $x \not\succ (x_{i,h}^\star, a^\star)$, $\phi^\star$ does not swap the recommended action (swap the recommended action to itself); (3) At any Type-I rechistory and Type-II rechistory with infoset $x \succ (x_{i,h}^\star, a^\star)$, $\phi^\star$ chooses action $\widehat{\pi}_i(x)$ (no matter seeing recommendation or not); (4) For any rechistory that does not fall into the above categories, $\phi^\star$ can be arbitrarily defined since those rechistories will not be encountered by the design of $\phi^\star$ as above. It is easy to see that such an 1-EFCE strategy modification $\phi^\star$ applied on any π_i implements the trigger policy $\mathsf{trig}(\pi_i, \widehat{\pi}_i, (x_{i,h}^\star, a^\star))$ so that their value functions are equal. This proves the claim.

Step 2. We next show that $\text{1-EFCEGap}(\overline{\pi}) \le \max_{i\in[m]} X_i A_i \cdot \text{TriggerGap}(\overline{\pi})$ for any correlated policy $\overline{\pi} \in \Delta(\Pi)$. For any 1-EFCE strategy modification $\phi \in \Phi_i^1$ and any $\pi_i \in \Pi_i$, by classifying x_h according to the first h such that $\phi \diamond \pi_i(x_i) \neq \pi_i(x_i)$, we have the decomposition of identity

$$\begin{aligned}1 = &\sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{a_h\in\mathcal{A}_i} \mathbf{1}\{x_h \text{ visited}, \ a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h\} \\ &+ \sum_{x_H\in\mathcal{X}_{i,H}} \sum_{a_H\in\mathcal{A}_i} \mathbf{1}\{x_H \text{ visited}, \ a_{1:H} \text{ recomd., and } \phi(x_H, a_{1:H}) = a_H\}.\end{aligned}$$

As a consequence, for any $\phi \in \Phi_i^1$, $\pi_i \in \Pi_i$ and $\pi_{-i} \in \Pi_{-i}$, we have

$$\begin{aligned}&V_i^{(\phi\diamond\pi_i)\times\pi_{-i}} - V_i^{\pi} = (\mathbb{E}_{(\phi\diamond\pi_i)\times\pi_{-i}} - \mathbb{E}_\pi)\left[\sum_{k=1}^{H} r_{i,k}\right] \\ &= (\mathbb{E}_{(\phi\diamond\pi_i)\times\pi_{-i}} - \mathbb{E}_\pi)\left[\sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{a_h\in\mathcal{A}_i} \mathbf{1}\{x_h \text{ visited}, \ a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h\} \sum_{k=1}^{H} r_{i,k}\right] \\ &\quad + (\mathbb{E}_{(\phi\diamond\pi_i)\times\pi_{-i}} - \mathbb{E}_\pi)\left[\sum_{x_H\in\mathcal{X}_{i,H}} \sum_{a_H\in\mathcal{A}_i} \mathbf{1}\{x_H \text{ visited}, \ a_{1:H} \text{ recomd., and } \phi(x_H, a_{1:H}) = a_H\} \sum_{k=1}^{H} r_{i,k}\right] \\ &= \sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{a_h\in\mathcal{A}_i} (\mathbb{E}_{(\phi\diamond\pi_i)\times\pi_{-i}} - \mathbb{E}_\pi)\left[\mathbf{1}\{x_h \text{ visited}, \ a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h\} \sum_{k=1}^{H} r_{i,k}\right],\end{aligned}$$

where the last equality used two facts: (i) fixing $x_1, a_1, \cdots, x_H, a_H$, supposing that $\phi(x_h, a_{1:h}) = a_h$ for all $h \le H$, then the probability of x_H is visited and $a_{1:H}$ are recommended are the same under $(\phi \diamond \pi_i) \times \pi_{-i}$ and π; (ii) the randomness of $\sum_{k=1}^{H} r_{i,k}$ is independent of policy when fixing $x_1, a_1, \cdots, x_H, a_H$. So the second quantity of left hand side of that equality is zero.

For any $(x_h, a_h) \in \mathcal{X}_{i,h} \times \mathcal{A}_i$, $\phi \in \Phi_i^1$ and $\pi_i \in \Pi_i$, we define the trigger policy $\mathsf{trig}(\pi_i, (\phi \diamond \pi_i), (x_h, a_h))$ to be a policy that plays π_i before triggered by (x_h, a_h) and plays $\phi \diamond \pi_i$ after triggered by (x_h, a_h). Supposing that $\phi(x_{h'}, a_{1:h'}) = a_{h'}, \forall h' < h$ and $\phi(x_h, a_{1:h}) \neq a_h$, the probability of x_h is visited and $a_{1:h}$ are recommended are the same the same under $(\phi \diamond \pi_i) \times \pi_{-i}$ and $\mathsf{trig}(\pi_i, (\phi \diamond \pi_i), (x_h, a_h)) \times \pi_{-i}$, which gives

$$\begin{aligned}&\mathbb{E}_{(\phi\diamond\pi_i)\times\pi_{-i}}\left[\mathbf{1}\{x_h \text{ visited}, \ a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h\} \sum_{h=1}^{H} r_{i,h}\right] \\ &= \mathbb{E}_{\mathsf{trig}(\pi_i,(\phi\diamond\pi_i),(x_h,a_h))\times\pi_{-i}}\left[\mathbf{1}\{x_h \text{ visited}, \ a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h\} \sum_{h=1}^{H} r_{i,h}\right].\end{aligned} \tag{18}$$

Consequently, we have

$$
\begin{aligned}
&\mathbb{E}_{\pi\sim\overline{\pi}}\Big(V_i^{(\phi\diamond\pi_i)\times\pi_{-i}} - V_i^{\pi}\Big)\\
&= \sum_{h=1}^{H}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{a_h\in\mathcal{A}_i}\mathbb{E}_{\pi\sim\overline{\pi}}\Big(\mathbb{E}_{(\phi\diamond\pi_i)\times\pi_{-i}} - \mathbb{E}_{\pi}\Big)\Big[\mathbf{1}\left\{x_h \text{ visited, } a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h\right\}\sum_{h=1}^{H} r_{i,h}\Big]\\
&\overset{(i)}{=} \sum_{h=1}^{H}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{a_h\in\mathcal{A}_i}\mathbf{1}\left\{\phi(x_h, a_{1:h}) \neq a_h\right\}\\
&\quad\times\mathbb{E}_{\pi\sim\overline{\pi}}\Big(\mathbb{E}_{\mathsf{trig}(\pi_i,(\phi\diamond\pi_i),(x_h,a_h))\times\pi_{-i}} - \mathbb{E}_{\pi}\Big)\Big[\mathbf{1}\left\{x_h \text{ visited } and\ a_{1:h} \text{ recomd.}\right\}\sum_{h=1}^{H} r_{i,h}\Big]\\
&\overset{(ii)}{=} \sum_{h=1}^{H}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{a_h\in\mathcal{A}_i}\mathbf{1}\left\{\phi(x_h, a_{1:h}) \neq a_h\right\}\mathbb{E}_{\pi\sim\overline{\pi}}\Big(\mathbb{E}_{\mathsf{trig}(\pi_i,(\phi\diamond\pi_i),(x_h,a_h))\times\pi_{-i}} - \mathbb{E}_{\pi}\Big)\Big[\sum_{h=1}^{H} r_{i,h}\Big]\\
&\overset{(iii)}{\leq} \sum_{h=1}^{H}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{a_h\in\mathcal{A}_i}\mathrm{TriggerGap}(\overline{\pi}) = X_iA_i\cdot\mathrm{TriggerGap}(\overline{\pi}).
\end{aligned}
$$

Here in (i) we used equation (18); in (iii) we bound the indicator by 1 and use the fact that TriggerGap is non-negative (by the observation that in the definition (17), we can choose x_i to be some leaf infoset $x_{i,H}$ and choose $\widehat{\pi}_i(x_{i,H}) = a$ so that $\mathsf{trig}(\pi_i, \widehat{\pi}_i, (x_i, a)) = \pi_i$); in (ii) we use the fact that $\mathsf{trig}(\pi_i, (\phi\diamond\pi_i), (x_h, a_h))$ and π are identical on any infoset x such that $x \neq x_h$ and $x \not\succ (x_h, a_h)$, so that

$$
\Big(\mathbb{E}_{\mathsf{trig}(\pi_i,(\phi\diamond\pi_i),(x_h,a_h))\times\pi_{-i}} - \mathbb{E}_{\pi}\Big)\Big[\mathbf{1}\left\{a_{1:h} \text{ are not recommended or } x_h \text{ is not visited}\right\}\sum_{h=1}^{H} r_{i,h}\Big] = 0.
$$

Finally, take supermum over $\phi \in \Phi_i^1$ and then take supermum over $i \in [m]$, we get

$$
1\text{-EFCEGap}(\overline{\pi}) = \max_{i\in[m]}\max_{\phi\in\Phi_i^1}\mathbb{E}_{\pi\sim\overline{\pi}}\Big(V_i^{(\phi\diamond\pi_i)\times\pi_{-i}} - V_i^{\pi}\Big) \leq \max_{i\in[m]} X_iA_i\cdot\mathrm{TriggerGap}(\overline{\pi}).
$$

This proves the lemma. □

C.2 Proof of Proposition C.2

Proof. We prove the containment result and strict containment result separately as follows.

Proof of K-EFCEGap$(\overline{\pi}) \leq (K+1)$-EFCEGap$(\overline{\pi})$ We claim that, for any $K \geq 0$ and strategy modification $\phi \in \Phi_i^K$, there exists $\phi' \in \Phi_i^{K+1}$ such that for any policy $\pi_i \in \Pi_i$, we have that $\phi \diamond \pi_i$ and $\phi' \diamond \pi_i$ gives the same policy. Given this claim, we have

$$
\max_{\phi\in\Phi_i^K}\mathbb{E}_{\pi\sim\overline{\pi}}V_i^{(\phi\diamond\pi_i)\times\pi_{-i}} \leq \max_{\phi\in\Phi_i^{K+1}}\mathbb{E}_{\pi\sim\overline{\pi}}V_i^{(\phi\diamond\pi_i)\times\pi_{-i}}.
$$

This implies K-EFCEGap$(\overline{\pi}) \leq (K+1)$-EFCEGap$(\overline{\pi})$ for all $K \geq 0$.

To prove such a claim, we can choose the strategy modification $\phi' \in \Phi_i^{K+1}$ to be the following: (1) For any rechistory $(x_{i,h}, b_{1:h-1}) \in \Omega_i^{(\mathrm{I}),K} \cap \Omega_i^{(\mathrm{I}),K+1}$ and any action $a_h \in \mathcal{A}_i$, we set $\phi'(x_{i,h}, b_{1:h-1}, a_h) = \phi(x_{i,h}, b_{1:h-1}, a_h)$; (2) For any rechistory $(x_{i,h}, b_{1:h-1}) \in \Omega_i^{(\mathrm{I}),K+1} \setminus \Omega_i^{(\mathrm{I}),K}$ and any action $a_h \in \mathcal{A}_i$, we set $\phi'(x_{i,h}, b_{1:h-1}, a_h) = \phi(x_{i,h}, b_{1:h_{\star\star}})$ where $h_{\star\star} = \inf\{k \leq h : \sum_{h''=1}^{k}\mathbf{1}\left\{a_{h''} \neq b_{h''}\right\} = K\}$; (3) For any rechistory $(x_{i,h}, b_{1:h'}) \in \Omega_i^{(\mathrm{II}),K+1}$, we set $\phi'(x_{i,h}, b_{1:h'}) = \phi(x_{i,h}, b_{1:h_{\star\star}})$ where $h_{\star\star} = \inf\{k \leq h' : \sum_{h''=1}^{k}\mathbf{1}\left\{a_{h''} \neq b_{h''}\right\} = K\}$. It is easy to see that for any $\pi_i \in \Pi_i$, we have that $\phi' \diamond \pi_i$ is the same as $\phi \diamond \pi_i$. This proves the claim.

Example of a game and a $\overline{\pi}$ **with** K-EFCEGap$(\overline{\pi}) = 0$ **but** $(K+1)$-EFCEGap$(\overline{\pi}) \geq 1/3$

For any $K \geq 0$, we consider a two-player game with $H = K+1$ steps and perfect information. The action spaces are $\mathcal{A}_1 = \{1, 2\}$ for the first player and $\mathcal{A}_2 = \{1, 2\}$ for the second player in each time step. The state space $\mathcal{S} = \cup_{h=1}^{H} \mathcal{S}_h$ can be identified as $\mathcal{S}_h = \mathcal{A}_1^{h-1} \times \mathcal{A}_2^{h-1}$ for $h = 1, 2, \dots, K+1$ and both players' infosets are the same as the state space $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{S}$. If action (a_h, b_h) is taken at $s_h = (a_{1:h-1}, b_{1:h-1}) \in \mathcal{S}_h$, then the environment will transit to the next state given by $s_{h+1} = (a_{1:h}, b_{1:h})$. The reward for the second player is always 0 at every time step. We design the reward for the first player (denoting as r_h in short) as following:

- The reward $r_h(\cdot, \cdot) = 0$ when $h \le K$ for every state actions.
- The reward at $s_{K+1} = (a_{1:K}, b_{1:K})$ is defined as

$$r_{K+1}(s_{K+1}, a_{K+1}, b_{K+1}) = \mathbf{1}\{a_1 \neq b_1, \dots, a_{K+1} \neq b_{K+1}\} + \frac{1}{2} \cdot \mathbf{1}\{a_1 = b_1, \dots, a_{K+1} = b_{K+1}\}.$$

Let $\Pi_\star = \{(\pi_\star, \pi_\star) : \pi_\star \in \Pi_1\}$ where Π_1 is the set of pure policies of the first player. That means, $\Pi_\star$ is the set of pure policies such that two players take the same action (either 1 or 2) at each state. We define $\overline{\pi}$ as the uniform distribution over such a policy space $\Pi_\star$. We claim that K-EFCEGap$(\overline{\pi}) = 0$ but $(K+1)$-EFCEGap$(\overline{\pi}) \ge 1/3$. Since the reward of the second player is always 0, we only need to consider the value function gap of the first player. Note that the value function of the first player for the correlated policy $\overline{\pi}$ is $1/2$.

We first consider the $(K+1)$-EFCEGap. If the first player deviates from the recommended action in every time step (this is an allowed strategy modification in Φ_i^{K+1}), she can receive reward 1 so that her received value is 1. As a consequence, we have

$$(K+1)\text{-EFCEGap}(\overline{\pi}) \ge 1 - 1/2 > 1/3.$$

We then consider the K-EFCEGap. If the first player chooses to deviate at any step, she need to play a different action from the second player at all time steps to receive an reward 1, otherwise she will receive reward 0. However, she is only allowed to see the recommendation K times. There is at least one time step such that she cannot see the recommendation and she need to guess what is the second player's action. The probability that her guess coincides with the other player's action is $1/2$ no matter how she guess. So by deviating from the recommended action, the first player can receive a value at most $1/2$. That means, K-EFCEGap$(\overline{\pi}) \le 1/2 - 1/2 = 0$. This finishes the proof of the proposition. $\square$

C.3 Proof of Proposition C.3

(a) We first show that ∞-EFCEGap$(\overline{\pi}) \le$ NFCEGap$(\overline{\pi})$. Indeed, for any ∞-EFCE strategy modification $\phi \in \Phi_i^{\infty\text{-EFCE}}$, we let $\phi' \in \Phi_i^{\text{NFCE}}$ such that for any $x_{i,h} \in \mathcal{X}_i$ and $\pi_i \in \Pi_i^{\text{pure}}$, $\phi'(x_{i,h}, \pi_i) := \phi(x_{i,h}, b_{1:h-1})$ where $b_{1:h-1} := (\pi_i(x_{i,1}), \dots, \pi_i(x_{i,h-1}))$ and $x_{i,1} \preceq \dots \prec x_{i,h-1} \prec x_{i,h}$ are the unique history of infosets leading to $x_{i,h}$. By comparing the execution of $\phi \diamond \pi$ (cf. Algorithm 1) and $\phi' \diamond \pi$ (cf. Definition C.1), the policy $\phi \diamond \pi_i$ exactly implements (i.e. is the same as) $\phi' \diamond \pi_i$. This gives

$$\infty\text{-EFCEGap}(\overline{\pi}) = \max_{\phi \in \Phi_i^{\infty\text{-EFCE}}} \mathbb{E}_{\pi \sim \overline{\pi}} V_i^{(\phi \diamond \pi_i) \times \pi_{-i}} \le \max_{\phi \in \Phi_i^{\text{NFCE}}} \mathbb{E}_{\pi \sim \overline{\pi}} V_i^{(\phi \diamond \pi_i) \times \pi_{-i}} = \text{NFCEGap}(\overline{\pi}).$$

We next prove the second claim (converse bound does not hold), by constructing the following example.

Example 1 (There exists an ∞-EFCE which is not ε-approximate NFCE with $\varepsilon = 1/20$)**:** We consider a two-player game with 2 steps and perfect information. The set of infosets $\mathcal{X}_{i,h}$ for both players gives $\mathcal{X}_{1,1} = \mathcal{X}_{2,1} = \{s_0\}$ and $\mathcal{X}_{1,2} = \mathcal{X}_{2,2} = \{s_{1,1}, s_{1,2}, s_{2,1}, s_{2,2}\}$. The action spaces are $\mathcal{A}_1 = \{a_1, a_2\}$ and $\mathcal{A}_2 = \{b_1, b_2\}$. If action pair (a_i, b_j) $(i, j = 1, 2)$ is chosen at s_0, $s_{i,j}$ would be reached with probability 1. The reward for the second player is always 0. And we design the reward for the first player as following:

- The reward at s_0 depends only on the action of the first player: the reward is $1/2$ if a_1 is chosen and 0 if a_2 is chosen.
- The rewards at $s_{1,1}$ and $s_{1,2}$ are always 0. The rewards at $s_{2,1}$ and $s_{2,2}$ depends only on the action of the second player: the rewards are all 1 if b_1 is chosen and 0 if b_2 is chosen.

Suppose $\overline{\pi}$ is the uniform distribution of all the deterministic policies that takes (a_1, b_1) or (a_2, b_2) at each infosets (there are $2^5 = 32$ such policies). We would verify that $\overline{\pi}$ is a ∞-EFCE but not a $1/20$-NFCE. Since the reward of the second player is always 0, we only need to consider the first player.

We first consider NFCE strategy modifications. On one hand, the first player only has motivation to modify his action at s_0 to a_2 if he observes that the recommendation at s_0, $s_{2,1}$ and $s_{2,2}$ are all a_1 (which happens iff the recommendation for his opponent are all b_1). Otherwise, he does not have motivation to change his action. If the first player choose such a modification, he will modify only $1/8$ of the deterministic policies, and for each deterministic policy π_i that are modified, the reward of the first player is increased by $1/2$. So using this NFCE strategy modification, the first player's value function are increased by $1/16$, which gives that

$$\mathrm{NFCEGap}(\overline{\pi}) \geq 1/16 > 1/20.$$

On the other hand, for any ∞-EFCE strategy modification, taking a_2 at s_0 always has utility $1/2$ since the actions taken by the second player at $\mathcal{X}_2$ are all uniformly distributed conditional on the recommendation at s_0. The utility of taking a_1 at s_0 is also $1/2$. This means that any ∞-EFCE strategy modification of $\overline{\pi}$ has value function $1/2$, so does $\overline{\pi}$. Consequently, $\overline{\pi}$ is an exact ∞-EFCE, i.e. ∞-EFCEGap$(\overline{\pi}) = 0$. $\diamond$

(b) Consider K-EFCE with $K = 0$. From the definition of strategy modifications and the executing of modified policy (Algorithm 1), for $\phi \in \Phi_i^0$ and policy $\pi_i \in \Pi_i^{\mathrm{pure}}$, $\phi \diamond \pi_i$ takes action $\phi(x_i, \emptyset)$ at x_i. So ϕ is equivalent to a modification $\phi' \in \Phi_i^{\mathrm{NFCCE}}$ which satisfies $\phi'(x_i) = \phi(x_i, \emptyset)$ for all $x_i \in \mathcal{X}_i$. Here, the equivalence means that $\phi \diamond \pi_i = \phi' \diamond \pi_i$ for any $\pi_i \in \Pi_i^{\mathrm{pure}}$. This gives

$$0\text{-EFCEGap}(\overline{\pi}) = \max_{\phi \in \Phi_i^0} \mathbb{E}_{\pi \sim \overline{\pi}} V_i^{(\phi \diamond \pi_i) \times \pi_{-i}} = \max_{\phi \in \Phi_i^{\mathrm{NFCCE}}} \mathbb{E}_{\pi \sim \overline{\pi}} V_i^{(\phi \diamond \pi_i) \times \pi_{-i}} = \mathrm{NFCCEGap}(\overline{\pi}),$$

which is the desired result. $\square$

D Properties of K-EFCE strategy modifications

For any $\phi \in \Phi_i^K$, we define its "probabilistic" expression μ^ϕ as follows: For any $a_h \in \mathcal{A}_i$,

$$\mu_h^\phi(a_h | x_{i,h}, b_{1:h-1}, b_h) := \mathbf{1}\{a_h = \phi(x_{i,h}, b_{1:h-1}, b_h)\} \quad \text{for all } (x_{i,h}, b_{1:h-1}), b_h \in \Omega_{i,h}^{(\mathrm{I}),K} \times \mathcal{A}_i,$$

$$\mu_h^\phi(a_h | x_{i,h}, b_{1:h'}) := \mathbf{1}\{a_h = \phi(x_{i,h}, b_{1:h'})\} \quad \text{for all } (x_{i,h}, b_{1:h'}) \in \Omega_{i,(h_\star,h)}^{(\mathrm{II}),K}.$$

In words, $\mu_h^\phi(\cdot | x_{i,h}, b_{1:h-1}, b_h) \in \Delta(\mathcal{A}_i)$ is the pure policy that takes action $\phi(x_{i,h}, b_{1:h-1}, b_h)$ deterministically, for any Type-I rechistory $(x_{i,h}, b_{1:h-1})$ and recommendation $b_h \in \mathcal{A}_i$; $\mu_h^\phi(a_h | x_{i,h}, b_{1:h'})$ is the pure policy that takes action $\phi(x_{i,h}, b_{1:h'})$ for any Type-II rechistory $(x_{i,h}, b_{1:h'})$. For convenience, we abuse notation slightly to let

$$\phi_h(\cdot | x_{i,h}, b_{1:h}) := \mu_h^\phi(\cdot | x_{i,h}, b_{1:h-1}, b_h), \quad \phi_h(\cdot | x_{i,h}, b_{1:h'}) := \mu_h^\phi(\cdot | x_{i,h}, b_{1:h'}).$$

Moreover, we use $\mathsf{D}(a_{1:k}, b_{1:k})$ to denote the Hamming distance of two action sequences $a_{1:k}, b_{1:k} \in \mathcal{A}_i^k$:

$$\mathsf{D}(a_{1:k}, b_{1:k}) := \sum_{h=1}^{k} \mathbf{1}\{a_h \neq b_h\},$$

and define the following notation as shorthand for the indicator that $a_{1:h}$ and $b_{1:h}$ differs in $\{\leq K-1, K\}$ elements:

$$\delta^{\leq K-1}(a_{1:h-1}, b_{1:h-1}) := \mathbf{1}\{\mathsf{D}(a_{1:h-1}, b_{1:h-1}) \leq K-1\};$$
$$\delta^{K}(a_{1:h-1}, b_{1:h-1}) := \mathbf{1}\{\mathsf{D}(a_{1:h-1}, b_{1:h-1}) = K\}.$$

Lemma D.1. *For any $\phi \in \Phi_i^K$ and any (potentially mixed) policy π_i for the i^{th} player, $\phi \diamond \pi_i$ is also a (potentially mixed) policy for the i^{th} player, with sequence-form expression*

$$(\phi \diamond \pi_i)_{1:h}(x_h, a_h) = \sum_{b_{1:h}} \delta^{\leq K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi_k(a_k | x_k, b_{1:k}) \prod_{k=1}^{h} \pi_i(b_k | x_k)$$

$$+\sum_{b_{1:h}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k\wedge\tau_K})\prod_{k=1}^{\tau_K}\pi_i(b_k|x_k),$$

where

$$\tau_K:=\inf\left\{h'\le h-1:\sum_{h''=1}^{h'}\mathbf{1}\left\{a_{h''}\neq b_{h''}\right\}\ge K\right\} \tag{19}$$

is the time step of the K-th deviation, and the event $\{\tau_K\le k\}$ can be determined by $(a_{1:k},b_{1:k})$ for any $k\ge 1$. Furthermore, we have

$$\sum_{(x_h,a_h)\in\mathcal{X}_{i,h}\times\mathcal{A}_i}\frac{(\phi\diamond\pi_i)_{1:h}(x_h,a_h)}{\pi^{\star,h}_{i,1:h}(x_h,a_h)}=X_{i,h}A_i.$$

Proof. Suppose the ancestors of x_h are $x_1\prec x_2\prec\cdots\prec x_{h-1}\prec x_h$ and the actions leading to x_h are $a_1,\dots,a_{h-1}$. The sequence-form expression $(\phi\diamond\pi_i)(x_h,a_h)$ is the probability of $(\phi\diamond\pi_i)$ choose a_k at x_k for all $k\in[h]$. We further denote the recommended action $b_k=\pi_i(x_k)$ for all $k\in[h]$.

If $|\{h'\in[h-1]:a_{h'}\neq b_{h'}\}|\le K-1$, the conditional probability of $(\phi\diamond\pi)$ choosing a_k at x_k for all $k\in[h]$ is $\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k})$, as the player would always swap the action; If $|\{h'\in[h]:a_{h'}\neq b_{h'}\}|\ge K$, the conditional probability of $(\phi\diamond\pi)$ choosing a_k at x_k for all $k\in[h]$ is $\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k\wedge\tau_K})$. So by the law of total probability, we have

$$\begin{aligned}(\phi\diamond\pi_i)_{1:h}(x_h,a_h)=&\sum_{b_{1:h}}\delta^{\le K-1}(a_{1:h-1},b_{1:h-1})\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k})\prod_{k=1}^h\pi_i(b_k|x_k)\\&+\sum_{b_{1:h}}\mathbf{1}\left\{|\{h'\in[h]:a_{h'}\neq b_{h'}\}|\ge K\right\}\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k\wedge\tau_K})\prod_{k=1}^h\pi_i(b_k|x_k).\end{aligned}$$

Notice that $\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k\wedge\tau_K})$ only depend on $b_{1:\tau_K}$ and $\sum_{b_{\tau_K:h}}\prod_{k=1}^h\pi_i(b_k|x_k)=\prod_{k=1}^{\tau_K}\pi_i(b_k|x_k)$, so the second summation admits a simpler form:

$$\begin{aligned}&\sum_{b_{1:h}}\mathbf{1}\left\{|\{h'\in[h]:a_{h'}\neq b_{h'}\}|\ge K\right\}\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k\wedge\tau_K})\prod_{k=1}^h\pi_i(b_k|x_k)\\=&\sum_{(h',b_{1:h'}):\sum_1^{h'}\mathbf{1}\{a_k\neq b_k\}=K\text{ and }b_{h'}\neq a_{h'}}\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k\wedge h'})\prod_{k=1}^{h'}\pi_i(b_k|x_k)\\=&\sum_{b_{1:h}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^h\phi_k(a_k|x_k,b_{1:k\wedge\tau_K})\prod_{k=1}^{\tau_K}\pi_i(b_k|x_k).\end{aligned}$$

The last equality is because we can append $a_{h'+1:h}$ to $b_{1:h'}$ to get a new $b_{1:h}$ which doesn't change the value of the summation. Then the first part of this lemma is proved. The second part actually is a direct corollary of Lemma B.4. □

The lemma above has the following corollary.

Corollary D.1. *For the i^{th} player and any pure policy $\pi\in\Pi$, fix any $\phi\in\Phi_i^K$ and $x_h\in\mathcal{X}_{i,h}$ with $(a_1,\dots,a_{h-1})$ being the unique history of actions leading to x_h. Then the probability that x_h is reached by the i^{th} player under policy $(\phi\diamond\pi_i)\times\pi_{-i}$ is*

$$\begin{aligned}&\mathbb{P}_{\phi\diamond\pi_i\times\pi_{-i}}\Big(x_h\text{ is reached by the }i^{th}\text{ player}\Big)\\=&\sum_{b_{1:h-1}}\delta^{\le K-1}(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k})\cdot\prod_{k=1}^{h-1}\pi_i(b_k|x_k)p_{1:h}^{\pi_{-i}}(x_{i,h})\end{aligned}$$

$$+\sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \cdot \prod_{k=1}^{\tau_K} \pi_i(b_k|x_k) p_{1:h}^{\pi_{-i}}(x_{i,h}).$$

Proof. We have

$$\begin{aligned}&\mathbb{P}_{\phi\diamond\pi_i\times\pi_{-i}}\Big(x_h \text{ is reached by the i}^{th} \text{ player}\Big)\\ &= \sum_{a\in\mathcal{A}_i} (\phi\diamond\pi_i)_{1:h}(x_h, a) p_{1:h}^{\pi_{-i}}(x_{i,h}),\end{aligned}$$

where $p_{1:h}^{\pi_{-i}}(x_{i,h})$ is defined in equation (12). So applying Lemma D.1 yields the desired result. □

As each step h, one (and only one) $x_h \in \mathcal{X}_{i,h}$ is visited, so the summation of the above reaching probability over x_h is 1. This directly yields the following corollary.

Corollary D.2. *For the i*th *player and any policy* $\pi \in \Pi$*, fix any* $\phi \in \Phi_i^K$*, we have*

$$\begin{aligned}&\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \cdot \prod_{k=1}^{h-1} \pi_i(b_k|x_k) p_{1:h}^{\pi_{-i}}(x_{i,h})\\ &+ \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \cdot \prod_{k=1}^{\tau_K} \pi_i(b_k|x_k) p_{1:h}^{\pi_{-i}}(x_{i,h}) = 1.\end{aligned}$$

E Regret decomposition for K-EFCE regret

This section presents the properties of K-EFCE regret, which will be useful for the proofs of our main results. Let $\{\pi^t\}_{t\in[T]}$ be a sequence of policies. Recall the K-EFCE regret for the i^{th} player (6):

$$R_{i,K}^T = \max_{\phi\in\Phi_i^K} \sum_{t=1}^T \Big(V_i^{\phi\diamond\pi_i^t\times\pi_{-i}^t} - V_i^{\pi^t}\Big). \tag{20}$$

Lemma E.1 (Online-to-batch for K-EFCE)**.** *Let* $\{\pi^t = (\pi_i^t)_{i\in[m]}\}_{t\in[T]}$ *be a sequence of product policies for all players over* T *rounds. Then, for the average (correlated) policy* $\overline{\pi} = \mathrm{Unif}(\{\pi^t\}_{t=1}^T)$*, we have*

$$K\text{-EFCEGap}(\overline{\pi}) = \max_{i\in[m]} R_{i,K}^T/T.$$

Proof. This follows directly by the definition of K-EFCEGap:

$$\begin{aligned}K\text{-EFCEGap}(\overline{\pi}) &= \max_{i\in[m]} \max_{\phi\in\Phi_i^K} \Big(V_i^{\phi\diamond\overline{\pi}} - V_i^{\overline{\pi}}\Big)\\ &= \max_{i\in[m]} \max_{\phi\in\Phi_i^K} \mathbb{E}_{\pi\sim\overline{\pi}}\Big[V_i^{\phi\diamond\pi_i\times\pi_{-i}} - V_i^{\pi}\Big]\\ &= \max_{i\in[m]} \max_{\phi\in\Phi_i^K} \frac{1}{T}\sum_{t=1}^T \Big[V_i^{\phi\diamond\pi_i^t\times\pi_{-i}^t} - V_i^{\pi^t}\Big]\\ &= \max_{i\in[m]} R_{i,K}^T/T.\end{aligned}$$

This proves the lemma. □

For $(x_{i,h}, b_{1:h-1}, \varphi) \in \Omega_i^{(\mathrm{I}),K} \times \Psi^s$, we define the immediate local swap regret as

$$\widehat{R}_{(x_h,b_{1:h-1}),\varphi}^T := \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big(\big\langle \pi_{i,h}^t(\cdot|x_{i,h}) - \varphi\diamond\pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot)\big\rangle\Big),$$

and the (overall) local swap regret as

$$\widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})} := \max_{\varphi} \widehat{R}^T_{(x_h,b_{1:h-1}),\varphi}. \tag{21}$$

For $(x_{i,h}, b_{1:h'}, a) \in \Omega_i^{(\mathrm{II}),K} \times \mathcal{A}_i$, we define the immediate local external regret as

$$\widehat{R}^T_{(x_h,b_{1:h'}),a} := \sum_{t=1}^T \prod_{k=1}^{h'} \pi_i^t(b_k|x_k)\Big(\big\langle \pi^t_{i,h}(\cdot|x_{i,h}), L^t_{i,h}(x_{i,h},\cdot)\big\rangle - L^t_{i,h}(x_{i,h},a)\Big),$$

and the (overall) local external regret as

$$\widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1:h'})} := \max_{a\in\mathcal{A}_i} \widehat{R}^T_{(x_h,b_{1:h'}),a}. \tag{22}$$

K-EFCE **regret decomposition** Our main result in this section is the following regret decomposition that decomposes the K-EFCE regret $R^T_{i,K}$ into combinations of local regrets at each rechistory.

Lemma E.2 (Regret decomposition for K-EFCE regret)**.** *We have $R^T_{i,K} \le \sum_{h=1}^H R^T_h$, with*

$$R^T_h := \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi) + \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h;\phi),$$

where

$$G_h^{T,\mathrm{swap}}(x_h;\phi) := \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k}) \widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})}, \tag{23}$$

and

$$\begin{aligned} G_h^{T,\mathrm{ext}}(x_h;\phi) &:= \sum_{b_{1:h-1}} \delta^{K}(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k\wedge\tau_K}) \widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1:\tau_K})} \\ &= \sum_{h'=K}^{h-1}\sum_{b_{1:h'}} \mathbf{1}\left\{\tau_K = h'\right\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k\wedge h'}) \widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1:h'})}. \end{aligned} \tag{24}$$

Above, $a_{1:h-1}$ is the unique sequence of actions leading to x_h, and τ_K (cf. definition in (19)) depends on $a_{1:h-1}$, $b_{1:h-1}$.

Proof of Lemma E.2. We begin by performing the following performance decomposition

$$\begin{aligned} &R^T_{i,K} = \max_{\phi\in\Phi_i^K} \sum_{t=1}^T (V_i^{\phi\diamond\pi_i^t\times\pi_{-i}^t} - V_i^{\pi^t}) \\ &= \max_{\phi\in\Phi_i^K} \sum_{t=1}^T \left(\mathbb{E}_{\phi\diamond\pi_i^t\times\pi_{-i}^t}\left[\sum_{h=1}^H r_{i,h}\right] - \mathbb{E}_{\pi^t}\left[\sum_{h=1}^H r_{i,h}\right]\right) \\ &= \max_{\phi\in\Phi_i^K} \sum_{t=1}^T\sum_{h=1}^H \left(\mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h}\pi^t_{i,h+1:H})\times\pi^t_{-i}}\left[\sum_{k=1}^H r_{i,k}\right] - \mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h-1}\pi^t_{i,h:H})\times\pi^t_{-i}}\left[\sum_{k=1}^H r_{i,k}\right]\right) \\ &= \max_{\phi\in\Phi_i^K} \sum_{t=1}^T\sum_{h=1}^H \left(\mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h}\pi^t_{i,h+1:H})\times\pi^t_{-i}}\left[\sum_{k=h}^H r_{i,k}\right] - \mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h-1}\pi^t_{i,h:H})\times\pi^t_{-i}}\left[\sum_{k=h}^H r_{i,k}\right]\right) \\ &= \max_{\phi\in\Phi_i^K} \sum_{h=1}^H\sum_{t=1}^T \left(\mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h-1}\pi^t_{i,h:H})\times\pi^t_{-i}}\left[\sum_{k=h}^H (1-r_{i,k})\right] - \mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h}\pi^t_{i,h+1:H})\times\pi^t_{-i}}\left[\sum_{k=h}^H (1-r_{i,k})\right]\right). \end{aligned}$$

Here, $((\phi\diamond\pi_i^t)_{1:h}\pi^t_{i,h+1:H})\times\pi^t_{-i}$ refers to the policy that the i^{th} player uses $\phi\diamond\pi_i^t$ for the first h step, and then uses π_i^t where as other players always use π^t_{-i}. The last step use the fact that $((\phi\diamond\pi_i^t)_{1:h}\pi^t_{i,h+1:H})\times\pi^t_{-i}$ and $((\phi\diamond\pi_i^t)_{1:h-1}\pi^t_{i,h:H})\times\pi^t_{-i}$ are the same for the first $h-1$ steps, so the expected reward in first $h-1$ steps are the same, too. Therefore, define

$$\widetilde{R}^T_h = \max_{\phi\in\Phi_i^K} \sum_{t=1}^T \left(\mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h-1}\pi^t_{i,h:H})\times\pi^t_{-i}}\left[\sum_{k=h}^H (1-r_{i,k})\right] - \mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h}\pi^t_{i,h+1:H})\times\pi^t_{-i}}\left[\sum_{k=h}^H (1-r_{i,k})\right]\right),$$

we have $R_{i,K}^T \le \sum_{h=1}^H \widetilde{R}_h^T$.

We next show that

$$\widetilde{R}_h^T \le \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi) + \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h;\phi) = R_h^T,$$

which yields the desired result. Fix any $h \in [H]$ and $\phi \in \Phi_i^K$, according to the execution of modified policy $\phi \diamond \pi_i$ as in Algorithm 1, we have

$$\begin{aligned}
&\mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h}\pi_{i,h+1:H}^t)\times\pi_{-i}^t}\left[\sum_{k=h}^H (1-r_{i,k})\right] \\
= &\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \underbrace{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K})}_{\text{probability of taking `right' actions leading to } x_h} \underbrace{\left\langle \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot)\right\rangle}_{\text{counterfactual loss}},
\end{aligned}$$

where we assume $\tau_K := \inf\{h' \le h-1 : \sum_{h''=1}^{h'} \mathbf{1}\{a_{h''} \ne b_{h''}\} \ge K\}$. Similarly,

$$\begin{aligned}
&\mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h-1}\pi_{i,h:H}^t)\times\pi_{-i}^t}\left[\sum_{k=h}^H (1-r_{i,k})\right] \\
= &\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \underbrace{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K})}_{\text{probability of taking `right' actions leading to } x_h} \underbrace{\left\langle \pi_i^t(\cdot|x_h), L_{i,h}^t(x_h,\cdot)\right\rangle}_{\text{counterfactual loss}}.
\end{aligned}$$

Substituting these into $\widetilde{R}_h^T$, we have

$$\begin{aligned}
\widetilde{R}_h^T &= \max_{\phi\in\Phi_i^K} \sum_{t=1}^T \left(\mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h-1}\pi_{i,h:H}^t)\times\pi_{-i}^t}\left[\sum_{k=h}^H (1-r_{i,k})\right] - \mathbb{E}_{((\phi\diamond\pi_i^t)_{1:h}\pi_{i,h+1:H}^t)\times\pi_{-i}^t}\left[\sum_{k=h}^H (1-r_{i,k})\right]\right) \\
&= \max_{\phi\in\Phi_i^K} \sum_{t=1}^T \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \left(\prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \left\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot)\right\rangle\right) \\
&= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \left(\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \left\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot)\right\rangle\right).
\end{aligned}$$

For fixed ϕ, x_h, based on whether $\mathsf{D}(a_{1:h-1}, b_{1:h-1}) \le K-1$ or not, we have

$$\begin{aligned}
\widetilde{R}_h^T &= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \left(\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \left\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot)\right\rangle\right) \\
&= \max_{\phi\in\Phi_i^K} (\mathrm{I}_h + \mathrm{II}_h),
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{I}_h := &\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\times \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \left\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot)\right\rangle,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{II}_h := &\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \mathbf{1}\{\mathsf{D}(a_{1:h-1}, b_{1:h-1}) \ge K\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \\
&\times \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \left\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot)\right\rangle.
\end{aligned}$$

For I_h, since all non-zero terms in the summation satisfy $\mathsf{D}(a_{1:h-1}, b_{1:h-1}) \le K-1$, at step h, the number of deviations is less than K, i.e. $\tau_K \ge h$. Moreover, max over $\phi \in \Phi_i^K$ can be separated into max over all $\phi_h(\cdot|x_h, b_{1:h\wedge\tau_K})$, so we have

$$\begin{aligned}
\max_{\phi\in\Phi_i^K} \mathrm{I}_h &= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \\
&\quad \times \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot) \Big\rangle \\
&\le \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\quad \times \max_{\varphi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_h) - (\varphi \diamond \pi_i^t)(\cdot|x_h), L_{i,h}^t(x_h,\cdot) \Big\rangle \\
&\overset{(i)}{=} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})} \\
&= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h).
\end{aligned}$$

Above, (i) follows by the definition of the local swap regret in (21). For II_h, since all non-zero terms in the summation satisfy $\mathsf{D}(a_{1:h-1}, b_{1:h-1}) \ge K$, all such $(a_{1:h-1}, b_{1:h-1})$ have already deviated K times at some step $h' \in [K, h-1]$, i.e. $\tau_K = h'$. In this case, the recommended action at x_h (i.e. $b_h \sim \pi_i^t(\cdot|x_h)$) cannot be observed. Thus for II_h, we have

$$\begin{aligned}
\max_{\phi\in\Phi_i^K} \mathrm{II}_h &= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \mathbf{1}\left\{\mathsf{D}(a_{1:h-1}, b_{1:h-1}) \ge K\right\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \\
&\quad \times \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge\tau_K}), L_{i,h}^t(x_h,\cdot) \Big\rangle \\
&= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \sum_{h'=K}^{h-1} \mathbf{1}\left\{\tau_K = h'\right\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \\
&\quad \times \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge h'}), L_{i,h}^t(x_h,\cdot) \Big\rangle \\
&= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{h'=K}^{h-1} \sum_{b_{1:h-1}} \mathbf{1}\left\{\tau_K = h'\right\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \\
&\quad \times \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge h'}), L_{i,h}^t(x_h,\cdot) \Big\rangle \\
&\overset{(i)}{=} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{h'=K}^{h-1} \sum_{b_{1:h'}} \mathbf{1}\left\{\tau_K = h'\right\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \\
&\quad \times \sum_{t=1}^{T} \prod_{k=1}^{h'} \pi_i^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h\wedge h'}), L_{i,h}^t(x_h,\cdot) \Big\rangle \\
&\le \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{h'=K}^{h-1} \sum_{b_{1:h'}} \mathbf{1}\left\{\tau_K = h'\right\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \\
&\quad \times \max_{a} \sum_{t=1}^{T} \prod_{k=1}^{h'} \pi_i^t(b_k|x_k) \big(L_{i,h}^t(x_h, \pi_i^t(x_h)) - L_{i,h}^t(x_h, a) \big) \\
&\overset{(ii)}{=} \max_{\phi\in\Phi_i^K} \sum_{h'=K}^{h-1} \sum_{b_{1:h'}} \mathbf{1}\left\{\tau_K = h'\right\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1:h'}),x_h}.
\end{aligned}$$

Here, (i) uses the fact that for fixed x_h, h' and $b_{1:h'}$ with $\tau_K = h'$, there exists only one $b_{h'+1:h-1}$ satisfying $\tau_K = h$ and $\sum_{b_{h'+1:h-1}} \prod_{k=h'+1}^{h-1} \pi_i^t(b_k|x_k) = 1$; (ii) follows by definition of the local external regret in (22). Furthermore, for fixed x_h, we can expand $b_{1:h'}$ to $b_{1:h} := (b_{1:h'}, a_{h^\star+1}, \cdots, a_{h-1})$ such that $\mathsf{D}(a_{1:h-1}, b_{1:h-1}) = K$. Then we can rewrite

$$
\begin{aligned}
&\sum_{h'=K}^{h-1} \sum_{b_{1:h'}} \mathbf{1}\{\tau_K = h'\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \widehat{R}^{T,\mathrm{ext}}_{(x_h, b_{1:h'}), x_h} \\
&= \sum_{h'=K}^{h-1} \sum_{b_{1:h-1}} \mathbf{1}\{\tau_K = h', \mathsf{D}(a_{1:h-1}, b_{1:h-1}) = K\} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \widehat{R}^{T,\mathrm{ext}}_{(x_h, b_{1:h'}), x_h} \\
&= \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge \tau_K}) \widehat{R}^{T,\mathrm{ext}}_{(x_h, b_{1:\tau_K}), x_h} = G_h^{T,\mathrm{ext}}(x_h).
\end{aligned}
$$

Consequently,

$$
\max_{\phi \in \Phi_i^K} \mathrm{II}_h \le \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h).
$$

Finally, combining the above bounds for $\max_{\phi \in \Phi_i^K} \mathrm{I}_h$ and $\max_{\phi \in \Phi_i^K} \mathrm{II}_h$ gives the desired result:

$$
\widetilde{R}_h^T = \max_{\phi \in \Phi_i^K} (\mathrm{I}_h + \mathrm{II}_h) \le \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h) + \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h) = R_h^T.
$$

□

F Proofs for Section 4

This section is devoted to proving Theorem 3.

The proof follows by bounding the K-EFCE regret (6):

$$
R_{i,K}^T = \max_{\phi \in \Phi_i^K} \sum_{t=1}^T \left(V_i^{\phi \diamond \pi_i^t \times \pi_{-i}^t} - V_i^{\pi^t} \right)
$$

for all players $i \in [m]$, and then converting to a bound on K-EFCEGap$(\overline{\pi})$ by the online-to-batch conversion (Lemma E.1).

By the regret decomposition for $R_{i,K}^T$ (Lemma E.2), we have $R_{i,K}^T \le \sum_{h=1}^H R_h^T$, where

$$
R_h^T := \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h; \phi) + \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h; \phi).
$$

The following two lemmas bound two terms

$$
\sum_{h=1}^H \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h; \phi) \quad \text{and} \quad \sum_{h=1}^H \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h; \phi).
$$

Their proofs are presented in Section F.1 & F.2 respectively.

Lemma F.1 (Bound on summation of $G_h^{T,\mathrm{swap}}(x_{i,h})$ with full feedback)**.** *If we choose learning rates as*

$$
\eta_{x_h} = \sqrt{\binom{H}{K \wedge H} X_i A_i^{K\wedge H} \log A_i / (H^2 T)}
$$

for all $x_h \in \mathcal{X}_{i,h}$ (same with (8)). Then we have

$$
\sum_{h=1}^H \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h; \phi) \le \sqrt{H^4 \binom{H}{K \wedge H} X_i A_i^{K\wedge H} T \log A_i}.
$$

Lemma F.2 (Bound on summation of $G_h^{T,\mathrm{ext}}(x_{i,h})$ with full feedback)**.** *If we choose learning rates as*

$$\eta_{x_h} = \sqrt{\binom{H}{K \wedge H} X_i A_i^{K\wedge h} \log A_i/(H^2T)}$$

for all $x_h \in \mathcal{X}_{i,h}$ (same with (8)). Then we have

$$\sum_{h=1}^{H} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h;\phi) \le \mathcal{O}\left(\sqrt{H^4\binom{H}{K\wedge H} X_i A_i^{K\wedge H} T \log A_i}\right).$$

Combining Lemma F.1 & F.2, we obtain regret bound

$$\begin{aligned} R_{i,K}^T &\le \sum_{h=1}^H R_h^T \le \sum_{h=1}^H \left(\max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi) + \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h;\phi)\right) \\ &= \mathcal{O}\left(\sqrt{H^4\binom{H}{K\wedge H} A_i^{K\wedge H} X_i T \log A_i}\right). \end{aligned} \tag{25}$$

In particular, as long as

$$T \ge \mathcal{O}\left(H^4 \binom{H}{K\wedge H}\left(\max_{i\in[m]} X_i A_i^{K\wedge H}\right)\log A_i/\varepsilon^2\right),$$

we have by the online-to-batch lemma (Lemma E.1) that the average policy $\overline{\pi} = \mathrm{Unif}(\{\pi^t\}_{t=1}^T)$ satisfies

$$K\text{-EFCEGap}(\overline{\pi}) = \frac{\max_{i\in[m]} R_{i,K}^T}{T} \le \max_{i\in[m]} \mathcal{O}\sqrt{H^4\binom{H}{K\wedge H}\left(\max_{i\in[m]} X_i A_i^{K\wedge H}\right)\log A_i/T} \le \varepsilon. \tag{26}$$

This proves Theorem 3. ☐

We remark that the above proof does not depend on the particular choice of π_{-i}^t, and thus the regret bound (25) also holds even if we control the i^{th} player only, and π_{-i}^t are arbitrary (potentially adversarial depending on all information before iteration t starts). This directly gives the following corollary.

Corollary F.1 (K-EFCE regret bound for K-EFR against adversarial opponents)**.** *For any $0 \le K \le \infty$, $\varepsilon \in (0, H]$, suppose the i^{th} player runs Algorithm 2 together against arbitrary (potentially adversarial) opponents $\{\pi_{-i}^t\}_{t=1}^T$, where* RegAlg *is instantiated as Algorithm 4 with learning rates specified in (8). Then the i^{th} player achieves K-*EFCE *regret bound:*

$$R_{i,K}^T = \max_{\phi\in\Phi_i^K} \sum_{t=1}^T \left(V_i^{\phi\diamond\pi_i^t\times\pi_{-i}^t} - V_i^{\pi^t}\right) \le \mathcal{O}\left(\sqrt{H^4\binom{H}{K\wedge H} X_i A_i^{K\wedge H} T\log A_i}\right).$$

F.1 Proof of Lemma F.1

Proof. Recall that $G_h^{T,\mathrm{swap}}(x_h;\phi)$ is defined as

$$G_h^{T,\mathrm{swap}}(x_h;\phi) := \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})},$$

where for each $h \in [H]$ and $(x_h, b_{1:h-1}) \in \Omega_i^{(\mathrm{I}),K}$,

$$\widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})} = \max_{\varphi} \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big(\big\langle \pi_{i,h}^t(\cdot|x_{i,h}) - \varphi\diamond\pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot)\big\rangle\Big).$$

For $x_h \in \mathcal{X}_{i,h}$, we first apply regret minimization lemma (Lemma A.2) on $\mathcal{R}_{x_h}$ to give an upper bound on $\widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})}$. Recall in Algorithm 2 with RegAlg instantiated as Algorithm 4, each $\mathcal{R}_{x_h}$

observes time selection functions $S^t_{b_{1:h-1}} = \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)$ for $(x_h, b_{1:h-1}) \in \Omega_i^{(\mathrm{I}),K}$ and loss vector $L^t_{i,h}(x_h, \cdot)$. Suppose $\mathcal{R}_{x_h}$ uses learning rate $\eta_{x_h} = \eta$ for all $x_h \in \mathcal{X}_i$. Then, by the regret bound with respect to a time selection function index and strategy modification pair (Lemma A.2), we have

$$\begin{aligned}
&\max_{\varphi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big(\big\langle \pi^t_{i,h}(\cdot|x_{i,h}) - \varphi \diamond \pi^t_{i,h}(\cdot|x_{i,h}), L^t_{i,h}(x_{i,h}, \cdot)\big\rangle \Big) \\
\le\ & \eta \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \|L^t_{i,h}(x_h, \cdot)\|_\infty \left\langle \pi_i^t(\cdot|x_h), L^t_{i,h}(x_h, \cdot)\right\rangle + \frac{(A_i + H)\log A_i}{\eta} \\
\le\ & \eta H^2 \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_h) + \frac{(A_i + H)\log A_i}{\eta}.
\end{aligned}$$

Above, we used (i) $\|L^t_{i,h}(x_h, \cdot)\|_\infty \le H p_{1:h}^{\pi^t_{-i}}(x_h)$, and (ii) $(|\mathcal{B}^s| + |\mathcal{B}^e|)|\Psi^s| \le A_i^{H+A_i}$ since the number of rechistories is no more than A_i^H .Then we can get

$$\begin{aligned}
&\max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h; \phi) \\
=\ & \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\times \max_{\varphi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big(L^t_{i,h}(x_h, \pi_i^t(x_h)) - L^t_{i,h}(x_h, \varphi \diamond \pi_i^t(x_h))\Big) \\
\le\ & \eta H^2 \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_h) \\
&+ \frac{(A_i + H)\log A_i}{\eta} \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}).
\end{aligned}$$

Letting

$$\mathrm{I}_h := \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k});$$

$$\mathrm{II}_h := \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_h).$$

For fixed $\phi \in \Phi_i^K$ and $x_h \in \mathcal{X}_{i,h}$, by counting the number of $b_{1:h-1}$ such that $\mathsf{D}(a_{1:h-1}, b_{1:h-1}) \le K-1$, we have

$$\begin{aligned}
&\sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
\le\ & \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \le \binom{h-1}{(K-1)\wedge(h-1)} A_i^{(K-1)\wedge(h-1)}.
\end{aligned}$$

Consequently,

$$\mathrm{I}_h \le X_{i,h} \binom{h-1}{(K-1)\wedge(h-1)} A_i^{K \wedge h-1}. \tag{27}$$

Note that by Corollary D.2, for fixed t,

$$\sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) \le 1.$$

Consequently, we have

$$\max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) \le T.$$

This yields that

$$\mathrm{II}_h \leq T. \tag{28}$$

Taking summation over $h \in [H]$, we have

$$\sum_{h=1}^{H} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi) \leq \sum_{h=1}^{H} \big(\frac{(A_i+H)\log A_i}{\eta}\mathrm{I}_h + H^2\eta\mathrm{II}_h\big)$$

$$\leq H^3\eta T + \sum_{h=1}^{H}\frac{(A_i+H)\log A_i}{\eta} X_{i,h}\binom{h-1}{(K-1)\wedge(h-1)} A_i^{K\wedge h-1}$$

$$\leq H^3\eta T + \frac{(A_i+H)\log A_i}{\eta} X_i\binom{H-1}{K\wedge H-1} A_i^{K\wedge H-1}$$

$$\leq H^3\eta T + \frac{H\log A_i}{\eta} X_i\binom{H}{K\wedge H} A_i^{K\wedge H}.$$

As we chose $\eta = \sqrt{\binom{H}{K\wedge H} X_i A_i^{K\wedge H}\log A_i/(H^2T)}$ per (8), we have

$$\sum_{h=1}^{H} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi) \leq \sqrt{H^4\binom{H}{K\wedge H} X_i A_i^{K\wedge H} T\log A_i}.$$

□

F.2 Proof of Lemma F.2

Proof. Recall that $G_h^{T,\mathrm{ext}}(x_h;\phi)$ is defined as

$$G_h^{T,\mathrm{ext}}(x_h;\phi) := \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge\tau_K})\widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1:\tau_K})}.$$

Since $(x_h, b_{1,\tau_K}) \in \Omega_i^{(\mathrm{II}),K}$, we have

$$\widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1,\tau_K})} = \max_a \sum_{t=1}^{T}\prod_{k=1}^{\tau_K}\pi_i^t(b_k|x_k)\Big(\big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot)\big\rangle - L_{i,h}^t(x_{i,h},a)\Big).$$

For $x_h \in \mathcal{X}_{i,h}$, we can apply regret minimization lemma (Lemma A.2) on $\mathcal{R}_{x_h}$ give an upper bound on $\widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})}$. Recall in Algorithm 2 with REGALG instantiated as Algorithm 4, each regret minimizer $\mathcal{R}_{x_h}$ observes time selection functions $S^t_{b_{1:h_{\tau_K}}} = \prod_{k=1}^{h_{\tau_K}}\pi_i^t(b_k|x_k)$ for $(x_h, b_{1:h_{\tau_K}}) \in \Omega_i^{(\mathrm{II}),K}$ and losses $L_{i,h}^t(x_h,\cdot)$. Suppose all $\mathcal{R}_{x_h}$ use the same learning rate η, by the bound on regret with respect to a time selection function index and strategy modification pair (Lemma A.2), we have

$$\max_a \sum_{t=1}^{T}\prod_{k=1}^{\tau_K}\pi_i^t(b_k|x_k)\Big(\big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot)\big\rangle - L_{i,h}^t(x_{i,h},a)\Big)$$

$$\leq\ \eta\sum_{t=1}^{T}\prod_{k=1}^{\tau_K}\pi_i^t(b_k|x_k)\|L_{i,h}^t(x_h,\cdot)\|_\infty\big\langle \pi_i^t(\cdot|x_h), L_{i,h}^t(x_h,\cdot)\big\rangle + \frac{\log((|\mathcal{B}^s|+|\mathcal{B}^e|)|\Psi^e|)}{\eta}$$

$$\leq\ \eta H^2\sum_{t=1}^{T}\prod_{k=1}^{\tau_K}\pi_i^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_h) + \frac{2H\log A_i}{\eta}.$$

Here, we use (i) $\|L_{i,h}^t(x_h,\cdot)\|_\infty \leq Hp_{1:h}^{\pi_{-i}^t}(x_h)$, and (ii) $(|\mathcal{B}^s|+|\mathcal{B}^e|)|\Psi^e| \leq A_i^{H+1}$ since the number of rechistories is no more than A_i^H . Then we can get

$$\max_{\phi\in\Phi_i^K} G_h^{T,\mathrm{ext}}(x_h;\phi)$$

$$= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K})$$
$$\times \max_a \sum_{t=1}^T \prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big(\big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot)\big\rangle - L_{i,h}^t(x_{i,h}, a)\Big)$$
$$\le \eta H^2 \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \cdot \sum_{t=1}^T \prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h)$$
$$+ \frac{2H\log A_i}{\eta} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}).$$

Similar to the proof of Lemma F.1, letting

$$\mathrm{I}_h := \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'});$$
$$\mathrm{II}_h := \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \cdot \sum_{t=1}^T \prod_{k=1}^{h'} \pi_i^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h).$$

For fixed $\phi \in \Phi_i^K$ and $x_h \in \mathcal{X}_{i,h}$, by counting the number of $b_{1:h-1}$ such that $\mathsf{D}(a_{1:h-1}, b_{1:h-1}) = K$, we have

$$\sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \le \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \le \binom{h-1}{K\wedge h} A_i^{K\wedge(h-1)}.$$

Consequently,

$$\mathrm{I}_h \le X_{i,h} \binom{h-1}{K\wedge h} A_i^{K\wedge(h-1)}. \tag{29}$$

Note that for fixed t, by Corollary D.2,

$$\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \prod_{k=1}^{h'} \pi_i^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \le 1.$$

Consequently, we have

$$\max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \sum_{t=1}^T \prod_{k=1}^{h'} \pi_i^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \le T.$$

This yields that

$$\mathrm{II}_h \le T. \tag{30}$$

Taking summation over $h \in [H]$, we have

$$\sum_{h=1}^H \max_{\phi\in\Phi_i^K} G_h^{T,\mathrm{ext}}(x_h;\phi) \le \sum_{h=1}^H \left(\frac{2H\log A_i}{\eta} \mathrm{I}_h + H^2\eta \mathrm{II}_h \right)$$
$$\le H^3\eta T + \sum_{h=1}^H \frac{2H\log A_i}{\eta} X_{i,h} \binom{h-1}{K\wedge h} A_i^{K\wedge H}$$
$$\le H^3\eta T + \frac{2H\log A_i}{\eta} X_i \binom{H-1}{K\wedge H} A_i^{K\wedge H}$$
$$\le H^3\eta T + \frac{2H\log A_i}{\eta} X_i \binom{H}{K\wedge H} A_i^{K\wedge H}.$$

As we choose $\eta = \sqrt{\binom{H}{K\wedge H} A_i^{K\wedge H} X_i \log A_i / (H^2 T)}$ per (8), we have

$$\sum_{h=1}^H \max_{\phi\in\Phi_i^K} G_h^{T,\mathrm{ext}}(x_h;\phi) \le 3\sqrt{H^4 \binom{H}{K\wedge H} X_i A_i^{K\wedge H} T \log A_i}.$$

□

Algorithm 6 Sample-based loss estimator for Type-I rechistories (i^{th} player's version)

Input: Policy π_i^t, π_{-i}^t. Balanced exploration policies $\{\pi_i^{\star,h}\}_{h\in[H]}$.

1: **for** $1 \le h \le H$, $\overline{W} \subseteq [h-1]$ with $|\overline{W}| = (K-1) \wedge (h-1)$ **do**

2: Set policy $\pi_i^{t,(h,\overline{W})} \leftarrow (\pi_{i,k}^{\star,h})_{k\in\overline{W}\cup\{h\}} \cdot (\pi_{i,k}^t)_{k\in[h-1]\setminus\overline{W}} \cdot \pi_{i,(h+1):H}^t$.

3: Play $\pi_i^{t,(h,\overline{W})} \times \pi_{-i}^t$ for one episode, observe trajectory

$$(x_{i,1}^{t,(h,\overline{W})}, a_{i,1}^{t,(h,\overline{W})}, r_{i,1}^{t,(h,\overline{W})}, \dots, x_{i,H}^{t,(h,\overline{W})}, a_{i,H}^{t,(h,\overline{W})}, r_{i,H}^{t,(h,\overline{W})}).$$

4: **for** all $(x_{i,h}, b_{1:h-1}) \in \Omega_i^{(\mathrm{I}),K}$ **do**

5: Find $(x_{i,1}, a_1) \prec \cdots \prec (x_{i,h-1}, a_{h-1}) \prec x_{i,h}$.

6: Set $\overline{W} \leftarrow \mathsf{fill}(\{k \in [h-1] : b_k \neq a_k\}, (K-1) \wedge (h-1))$.

7: Construct loss estimator for all $a \in \mathcal{A}_i$

$$\widetilde{L}_{(x_{i,h},b_{1:h-1})}^t(a) \leftarrow \frac{\mathbf{1}\left\{(x_{i,h}^{t,(h,\overline{W})}, a_{i,h}^{t,(h,\overline{W})}) = (x_{i,h}, a)\right\}}{\pi_{i,1:h}^{t,(h,\overline{W})}(x_{i,h}, a)} \cdot \sum_{h''=h}^{H} \left(1 - r_{i,h''}^{t,(h,\overline{W})}\right). \tag{32}$$

Output: Loss estimators $\left\{\widetilde{L}_{(x_{i,h},b_{1:h-1})}^t(\cdot)\right\}_{(x_{i,h},b_{1:h-1})\in\Omega_i^{(\mathrm{I}),K}}$.

G Proofs for Section 5

This section is devoted to proving Theorem 5. The additional notation presented at the beginning of Section F is also used in this section.

G.1 Sample-based loss estimator for Type-I rechistories

We first present the sample-based loss estimator for Type-I rechistories in Algorithm 6, complementary to the Type-II case presented in the main text (Algorithm 3). Here, Line 6 uses the "fill operator" defined as follows: For any index set $I \subset \mathbb{Z}_{\ge 1}$ and $n' \ge |I|$, $\mathsf{fill}(I, n')$ is defined as the unique superset of I with size n' and the smallest possible additional elements, i.e.

$$\mathsf{fill}(I, n') := I \cup \{I_{(1)}^c, \dots, I_{(n'-|I|)}^c\}, \tag{31}$$

where $I^c := \mathbb{Z}_{\ge 1} \setminus I$ with sorted elements $I_{(1)}^c < I_{(2)}^c < \cdots$. For example, $\mathsf{fill}(\{1,3,8\}, 5) = \{1,3,8\} \cup \{2,4\} = \{1,2,3,4,8\}$.

G.2 Algorithm description of Balanced K-EFR

Algorithm 7 presents the detailed description of Balanced K-EFR.

For each infoset $x_{i,h} \in \mathcal{X}_i$, the algorithm computes time selection functions $\{S_{\mathbf{b}}^t\}_{\mathbf{b}\in\Omega_i^{(\mathrm{I}),K}(x_{i,h})\cup\Omega_i^{(\mathrm{II}),K}(x_{i,h})}$, based on the corresponding M^t (defined in Line 5 & 6), as well as the following additional weighting function (below $W := \{k \in [h-1] : a_k \neq b_k\}$, and $\mathsf{fill}(\cdot,\cdot)$ is defined in (31))

$$w_{b_{1:h-1}}(x_{i,h}) = \prod_{k\in\mathsf{fill}(W,(h-1)\wedge(K-1))\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k), \quad \text{for all } b_{1:h-1} \in \Omega_i^{(\mathrm{I}),K}(x_{i,h}), \tag{33}$$

$$w_{b_{1:h'}}(x_{i,h}) = \prod_{k\in W\cup\{h'+1,\cdots,h\}} \pi_{i,k}^{\star,h}(a_k|x_k), \quad \text{for all } b_{1:h'} \in \Omega_i^{(\mathrm{II}),K}(x_{i,h}). \tag{34}$$

The resulting choice of time selection functions, $S_{\mathbf{b}}^t = M_{\mathbf{b}}^t w_{\mathbf{b}}^t(x_{i,h})$, is different from Algorithm 2, and is needed for this sampled case.

Self-play protocol Here we explain the protocol of how we let all players play Algorithm 7 for T rounds via self-play in Theorem 5. Within each round, each player first determines her own policy π_i^t by Line 3-8. Then, we let all players compute their sample-based loss estimators in a round-robin fashion: The first player obtains her loss estimators first (Line 9) by playing the sampling policies within Algorithm 6 & 3, in which case all other players keep playing π_{-1}^t. Then the same procedure goes on for players $2, \dots, m$. Note that overall, each round plays m times the number of episodes required for each player (specified by Algorithm 6 & 3). The following Lemma gives a bound on this number of episodes.

Lemma G.1 (Number of episodes played by sampling algorithms)**.** *One call of Algorithm 6 and Algorithm 3 plays (combinedly) $\binom{H+1}{K\wedge H+1} + K \wedge H - 1 \le 3H\binom{H}{K\wedge H}$ episodes.*

Proof. The proof follows by counting the number of $\overline{W}$'s and W's in the sampling algorithms, i.e. the cardinalities of

$$\mathcal{W}_{1,h} := \left\{\overline{W} : \overline{W} \subseteq [h-1] \text{ with } |\overline{W}| = (K-1) \wedge (h-1)\right\}$$

for $1 \le h \le H$, which comes from Line 1 in Algorithm 6, and

$$\mathcal{W}_{2,h',h} := \{W : K \le h' < h, W \subseteq [h'] \text{ with } |W| = K \text{ and ending in } h'\}$$

for $K \le h \le H$, which comes from line 1 in Algorithm 3.

For the first kind of sets , we have

$$\begin{aligned}
\sum_{h=1}^{H} |\mathcal{W}_{1,h}| &= \sum_{h=1}^{K\wedge H-1} \binom{h-1}{(K-1)\wedge(h-1)} + \sum_{h=K\wedge H}^{H} \binom{h-1}{(K-1)\wedge(h-1)} \\
&= K \wedge H - 1 + \sum_{h=K}^{H} \binom{h-1}{K\wedge H-1} \\
&= K \wedge H - 1 + \binom{H}{K \wedge H}.
\end{aligned}$$

For the second kind of sets , we have

$$\begin{aligned}
\sum_{K\le h'<h\le H} |\mathcal{W}_{2,h',h}| &= \sum_{K\le h'<h\le H} \binom{h'-1}{K-1} \\
&= \sum_{h=K}^{H}\sum_{h'=K}^{h-1} \binom{h'-1}{K-1} = \sum_{h=K}^{H} \binom{h-1}{K} \\
&= \binom{H}{K+1} = \binom{H}{K\wedge H+1}.
\end{aligned}$$

Taking summation gives that the number of episodes equals

$$\sum_{h=1}^{H} |\mathcal{W}_{1,h}| + \sum_{K\le h'<h\le H} |\mathcal{W}_{2,h',h}| = \binom{H+1}{K\wedge H+1} + K \wedge H - 1.$$

Finally, we show the above quantity can be upper bounded by $3H\binom{H}{K\wedge H}$. For $K \ge H$, we have $K \wedge H = H$ and the above quantity is $1 + H - 1 = H \le 3H = 3H\binom{H}{K\wedge H}$. For $K < H$, we have $K \wedge H = K$, and thus

$$\begin{aligned}
&\binom{H+1}{K\wedge H+1} + K \wedge H - 1 = \binom{H+1}{K+1} + K - 1 = \binom{H}{K} \cdot \frac{H+1}{K+1} + K - 1 \\
&\le 2H \cdot \binom{H}{K} + H \le 3H \cdot \binom{H}{K},
\end{aligned}$$

where the last inequality follows from the fact that $\binom{H}{K} \ge 1$. This is the desired bound. □

Algorithm 7 Balanced K-EFR (i^{th} player's version)

Input: Weights $\{w_{b_{1:h-1}}(x_{i,h})\}_{x_{i,h},b_{1:h-1}\in\Omega_i^{(\mathrm{I}),K}(x_{i,h})}$ and $\{w_{b_{1:h'}}(x_{i,h})\}_{x_{i,h},b_{1:h'}\in\Omega_i^{(\mathrm{II}),K}(x_{i,h})}$ defined in (33), (34), learning rates $\{\eta_{x_{i,h}}\}_{x_{i,h}\in\mathcal{X}_i}$, loss upper bound $\overline{L}>0$.

Initialize regret minimizers $\{\mathcal{R}_{x_{i,h}}\}_{x_{i,h}\in\mathcal{X}_i}$ with REGALG, learning rate $\eta_{x_{i,h}}$, and loss upper bound $\overline{L}$.

for iteration $t=1,\dots,T$ **do**

  for $h=1,\dots,H$ **do**

    for $x_{i,h}\in\mathcal{X}_{i,h}$ **do**

      $S^t_{b_{1:h-1}} := M^t_{b_{1:h-1}} w_{b_{1:h-1}}(x_{i,h})$ where $M^t_{b_{1:h-1}} := \prod_{k=1}^{h-1}\pi^t_{i,k}(b_k|x_k)$.

      $S^t_{b_{1:h'}} := M^t_{b_{1:h'}} w_{b_{1:h'}}(x_{i,h})$ where $M^t_{b_{1:h'}} := \prod_{k=1}^{h'}\pi^t_{i,k}(b_k|x_k)$.

      $\mathcal{R}_{x_{i,h}}$.OBSERVE_TIMESELECTION$(\{S^t_{b_{1:h-1}}\}_{b_{1:h-1}\in\Omega_i^{(\mathrm{I}),K}(x_{i,h})}\cup\{S^t_{b_{1:h'}}\}_{b_{1:h'}\in\Omega_i^{(\mathrm{II}),K}(x_{i,h})})$.

      Set policy $\pi_i^t(\cdot|x_{i,h})\leftarrow\mathcal{R}_{x_{i,h}}$.RECOMMEND$()$.

  Obtaining sample-based loss estimators

$$\left\{\widetilde{L}^t_{(x_{i,h},b_{1:h-1})}\right\}_{b_{1:h-1}\in\Omega_i^{(\mathrm{I}),K}(x_{i,h})}\quad\text{and}\quad\left\{\widetilde{L}^t_{(x_{i,h},b_{1:h'})}\right\}_{b_{1:h'}\in\Omega_i^{(\mathrm{II}),K}(x_{i,h})}$$

  from Algorithm 3 & 6 respectively.

  for all $x_{i,h}\in\mathcal{X}_i$ **do**

    $\mathcal{R}_{x_{i,h}}$.OBSERVE_LOSS$(\{\widetilde{L}^t_{(x_{i,h},b_{1:h-1})}\}_{b_{1:h-1}\in\Omega_i^{(\mathrm{I}),K}(x_{i,h})}\cup\{\widetilde{L}^t_{(x_{i,h},b_{1:h'})}\}_{b_{1:h'}\in\Omega_i^{(\mathrm{II}),K}(x_{i,h})})$.

Output: Policies $\{\pi_i^t\}_{t\in[T]}$.

G.3 Proof of Theorem 5

The proof follows a similar structure as the proof of Theorem 3 (cf. Section F), with different bounds on the regret terms and bounds on additional concentration terms.

Proof. The proof follows by bounding the K-EFCE regret (6):

$$R^T_{i,K}=\max_{\phi\in\Phi_i^K}\sum_{t=1}^T\left(V_i^{\phi\diamond\pi_i^t\times\pi_{-i}^t}-V_i^{\pi^t}\right)$$

for all players $i\in[m]$, and then converting to a bound on K-EFCEGap$(\overline{\pi})$ by the online-to-batch conversion (Lemma E.1).

By the regret decomposition for $R^T_{i,K}$ (Lemma E.2), we have $R^T_{i,K}\le\sum_{h=1}^H R^T_h$, where

$$R^T_h:=\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\text{swap}}(x_h;\phi)+\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\text{ext}}(x_h;\phi).$$

We bound the terms

$$\sum_{h=1}^H\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\text{swap}}(x_h;\phi)\quad\text{and}\quad\sum_{h=1}^H\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\text{ext}}(x_h;\phi)$$

when we play Balanced K-EFR (Algorithm 7) in the following two lemmas. Their proofs are presented in Section G.4 & G.5 respectively.

Lemma G.2 (Bound on summation of $G_h^{T,\text{swap}}(x_{i,h})$ with bandit feedback)**.** *If we choose learning rates as*

$$\eta_{x_h}=\sqrt{\binom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\iota/(H^3T)}$$

for all $x_h\in\mathcal{X}_i$ (same with (10)). With probability at least $1-p/2$, we have

$$\sum_{h=1}^H\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\text{swap}}(x_h;\phi)\le\mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_iT\iota}\right)+\mathcal{O}\left(H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\right)$$

$$+\mathcal{O}\left(\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\sqrt{\frac{H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota}{T}}\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

Lemma G.3 (Bound on summation of $G_h^{T,\mathrm{ext}}(x_{i,h})$ with bandit feedback)**.** *If we choose learning rates as*

$$\eta_{x_h} = \sqrt{\binom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\iota/(H^3T)}$$

for all $x_h \in \mathcal{X}_i$ (same with (10)). With probability at least $1-p/2$, we have

$$\sum_{h=1}^{H}\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\mathrm{ext}}(x_h;\phi) \le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_iT\iota}\right)+\mathcal{O}\left(H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\right)$$
$$+\mathcal{O}\left(\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\sqrt{\frac{H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota}{T}}\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

By Lemma G.2, G.3, and a union bound for all $i\in[m]$, we get

$$\begin{aligned}
R_{i,K}^T &\le \sum_{h=1}^{H} R_h^T \\
&\le \sum_{h=1}^{H}\left(\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\mathrm{swap}}(x_h;\phi)+\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}G_h^{T,\mathrm{ext}}(x_h;\phi)\right) \\
&\le \mathcal{O}\Bigg(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_iT\iota}+H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota \\
&\qquad +\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\sqrt{\frac{H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota}{T}}\Bigg).
\end{aligned}$$

with probability at least $1-p$ for all $i\in[m]$ simultaneously, where $\iota = \log(8\sum_{j\in[m]}X_jA_j/p)$.

Further using the "trivial" bound $R_{i,K}^T \le HT$ (by the fact that $V_i^\pi\in[0,H]$ for any joint policy π) gives

$$\begin{aligned}
R_{i,K}^T &\overset{(i)}{\le} HT\cdot\min\left\{1,\mathcal{O}\left(\sqrt{H\binom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\iota/T}\right)\right\} \\
&\le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_iT\iota}\right),
\end{aligned}$$

where (i) follows by noticing that:

- if $T < H\binom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\iota$, $R_{i,K}^T \le HT = HT\min\left\{1,\mathcal{O}\left(\sqrt{H\binom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\iota/T}\right)\right\}$;
- if $T \ge H\binom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\iota$, $R_{i,K}^T \le HT\cdot\mathcal{O}\left(\sqrt{H\binom{H}{K\wedge H}X_iA_i^{K\wedge H+1}\iota/T}\right)$.

Therefore, as long as

$$T \ge \mathcal{O}\left(H^3\binom{H}{K\wedge H}\left(\max_{i\in[m]}X_iA_i^{K\wedge H+1}\right)\iota/\varepsilon^2\right),$$

we have by the online-to-batch lemma (Lemma E.1) that the average policy $\overline{\pi} = \mathrm{Unif}(\{\pi^t\}_{t=1}^T)$ satisfies

$$K\text{-EFCEGap}(\overline{\pi}) = \frac{\max_{i\in[m]} R_{i,K}^T}{T} \le \max_{i\in[m]} \mathcal{O}\sqrt{H^3\binom{H}{K\wedge H}\left(\max_{i\in[m]} X_i A_i^{K\wedge H+1}\right)\iota/T} \le \varepsilon.$$

This proves the first part of Theorem 5.

Finally, we count how many episodes are played at each iteration. By our self-play protocol (cf. Section G.2) and Lemma G.1, each iteration involves m rounds of sampling (one for each player), where each round plays at most $3H\binom{H}{K\wedge H}$ episodes. Therefore, each iteration plays at most $3mH\binom{H}{K\wedge H}$ episodes, and so the total number of episodes played by Algorithm 7 is

$$3mH\binom{H}{K\wedge H}\cdot T = \mathcal{O}\left(mH^4\binom{H}{K\wedge H}^2\left(\max_{i\in[m]} X_i A_i^{K\wedge H+1}\right)\iota/\varepsilon^2\right).$$

This is the desired result. □

G.4 Proof of Lemma G.2

Proof. Recall that $G_h^{T,\mathrm{swap}}(x_h;\phi)$ is defined as (eq. (21))

$$G_h^{T,\mathrm{swap}}(x_h;\phi) := \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})}.$$

For each $h\in[H]$ and $(x_h, b_{1:h-1}) \in \Omega_i^{(\mathrm{I}),K}$, we have

$$\begin{aligned}
\widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})} &= \max_{\varphi} \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big(\big\langle \pi_{i,h}^t(\cdot|x_{i,h}) - \varphi\diamond\pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot)\big\rangle \Big) \\
&\le \max_{\varphi} \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}) - \varphi\diamond\pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:h-1})}\Big\rangle \\
&\quad + \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot) - \widetilde{L}^t_{(x_h,b_{1:h-1})}\Big\rangle \\
&\quad + \max_{\varphi} \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big\langle \varphi\diamond\pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:h-1})} - L_{i,h}^t(x_{i,h},\cdot)\Big\rangle.
\end{aligned}$$

Substituting this into $\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi)$ yields that

$$\begin{aligned}
\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi) = &\max_{\phi\in\Phi_i^K}\sum_{x_h}\sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \widehat{R}^{T,\mathrm{swap}}_{(x_h,b_{1:h-1})} \\
\le\ & \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} + \mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} + \mathrm{BIAS}_{2,h}^{T,\mathrm{swap}},
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} := &\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\times \max_{\varphi} \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}) - \varphi\diamond\pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:h-1})}\Big\rangle, \\
\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} := &\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\times \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k)\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot) - \widetilde{L}^t_{(x_h,b_{1:h-1})}\Big\rangle
\end{aligned}$$

$$\mathrm{BIAS}_{2,h}^{T,\mathrm{swap}} := \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}),$$
$$\times \max_{\varphi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \varphi \diamond \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:h-1})} - L_{i,h}^t(x_{i,h},\cdot) \Big\rangle.$$

The bounds for $\sum_{h=1}^H \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}}$, $\sum_{h=1}^H \mathrm{BIAS}_{1,h}^{T,\mathrm{swap}}$, and $\sum_{h=1}^H \mathrm{BIAS}_{2,h}^{T,\mathrm{swap}}$ are given in the following three lemmas (proofs deferred to Appendix G.4.1, G.4.2, and G.4.3) respectively.

Lemma G.4 (Bound on $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}}$). *If we choose learning rates as*

$$\eta_{x_h} = \sqrt{\binom{H}{K\wedge H} X_i A_i^{K\wedge H+1} \log(X_i A_i/p)/(H^3 T)}\cdot$$

for all $x_h \in \mathcal{X}_i$ (same with (10)). Then with probability at least $1-p/4$, we have

$$\sum_{h=1}^H \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} \le \sqrt{H^3 \binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i T\iota}$$
$$+\mathcal{O}\left(\binom{H}{K\wedge H} A_i^{K\wedge H} X_i \iota \sqrt{\frac{H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota}{T}}\right)$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

Lemma G.5 (Bound on $\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}}$). *With probability at least $1-\frac{p}{8}$, we have*

$$\sum_{h=1}^H \mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} \le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H} A_i^{K\wedge H} X_i T\iota} + H\binom{H}{K\wedge H-1} A_i^{K\wedge H} X_i\iota\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

Lemma G.6 (Bound on $\mathrm{BIAS}_{2,h}^{T,\mathrm{swap}}$). *With probability at least $1-\frac{p}{8}$, we have*

$$\sum_{h=1}^H \mathrm{BIAS}_{2,h}^{T,\mathrm{swap}} \le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i T\iota} + H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

Combining Lemma G.4, G.5, and G.6, we have with probability at least $1-p/2$ that

$$\sum_{h=1}^H \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{swap}}(x_h;\phi) \le \sum_{h=1}^H \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} + \sum_{h=1}^H \mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} + \sum_{h=1}^H \mathrm{BIAS}_{2,h}^{T,\mathrm{swap}}$$
$$\le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i T\iota}\right)$$
$$+\mathcal{O}\left(\binom{H}{K\wedge H} A_i^{K\wedge H} X_i\iota\sqrt{\frac{H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota}{T}}\right)$$
$$+\mathcal{O}\left(H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota\right).$$

□

G.4.1 Proof of Lemma G.4: Bound on $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}}$

Proof. Recall that $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}}$ is defined as

$$\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^{\le K-1}(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k})$$
$$\times\max_{\varphi}\sum_{t=1}^{T}\prod_{k=1}^{h-1}\pi_i^t(b_k|x_k)\Big\langle\pi_i^t(\cdot|x_h)-\varphi\diamond\pi_i^t(\cdot|x_h),\widetilde{L}_{(x_h,b_{1:h-1})}^t\Big\rangle.$$

We first apply regret minimization lemma (Lemma A.3) on $\mathcal{R}_{x_h}$ to give an upper bound of

$$\max_{\varphi}\sum_{t=1}^{T}\prod_{k=1}^{h-1}\pi_i^t(b_k|x_k)\Big\langle\pi_i^t(\cdot|x_h)-\varphi\diamond\pi_i^t(\cdot|x_h),\widetilde{L}_{(x_h,b_{1:h-1})}^t\Big\rangle.$$

Recall that in Algorithm 7, each regret minimizer $\mathcal{R}_{x_h}$ is associated with weights $M_{b_{1:h-1}}^t=\prod_{k=1}^{h-1}\pi_i^t(b_k|x_k)$ and $w_{b_{1:h-1}}=\prod_{k\in\mathsf{fill}(W,(h-1)\wedge(K-1))\cup\{h\}}\pi_{i,k}^{\star,h}(a_k|x_k)$ for any $b_{1:h-1}\in\Omega_i^{(\mathrm{I}),K}(x_h)$, where $W=\{k\in[h-1]:b_k\neq a_k\}$. Let $\overline{W}=\mathsf{fill}(W,(h-1)\wedge(K-1))$, $\eta_{x_h}=\eta$ be the learning rate of $\mathcal{R}_{x_h}$ and $\overline{L}=H$. Since regret minimizers $\mathcal{R}_{x_h}$ observe $\Big\{\widetilde{L}_{(x_{i,h},b_{1:h'})}^t(\cdot)\Big\}_{(x_{i,h},b_{1:h'})\in\Omega_i^{(\mathrm{II}),K}}$ and $\Big\{\widetilde{L}_{(x_{i,h},b_{1:h-1})}^t(\cdot)\Big\}_{(x_{i,h},b_{1:h-1})\in\Omega_i^{(\mathrm{I}),K}}$ from Algorithm 3 & 6 as its loss vector at round t, we have

$$\begin{aligned}
&M_{b_{1:h-1}}^t w_{b_{1:h-1}}\widetilde{L}_{(x_{i,h},b_{1:h-1})}^t(\cdot)\\
&=M_{b_{1:h-1}}^t w_{b_{1:h-1}}\frac{\mathbf{1}\Big\{(x_{i,h}^{t,(h,\overline{W})},a_{i,h}^{t,(h,\overline{W})})=(x_{i,h},\cdot)\Big\}}{\prod_{k\in\overline{W}\cup\{h\}}\pi_{i,k}^{\star,h}(a_k|x_{i,k})\prod_{k\in[h-1]\setminus\overline{W}}\pi_i^t(b_k|x_{i,k})}\cdot\sum_{h''=h}^{H}\Big(1-r_{i,h''}^{t,(h,\overline{W})}\Big)\\
&=\prod_{k\in\overline{W}}\pi_i^t(b_k|x_{i,k})\mathbf{1}\Big\{(x_{i,h}^{t,(h,\overline{W})},a_{i,h}^{t,(h,\overline{W})})=(x_{i,h},\cdot)\Big\}\sum_{h''=h}^{H}\Big(1-r_{i,h''}^{t,(h,\overline{W})}\Big)\\
&\in[0,H]=[0,\overline{L}].
\end{aligned}$$

Similarly, we have $M_{b_{1:h'}}^t w_{b_{1:h-1}}\widetilde{L}_{(x_{i,h},b_{1:h'})}(\cdot)\in[0,\overline{L}]$. Moreover, let $\mathcal{F}_{t-1}$ be the σ-algebra containing all the information until π^t is sampled, by the sampling algorithm, we have that the loss estimators are unbiased:

$$\mathbb{E}\Big[\widetilde{L}_{(x_{i,h},b_{1:h-1})}^t(\cdot)|\mathcal{F}_{t-1}\Big]=\mathbb{E}\Big[\widetilde{L}_{(x_{i,h},b_{1:h'})}^t(\cdot)|\mathcal{F}_{t-1}\Big]=L_h^t(x_{i,h},\cdot),$$

for all $(x_{i,h},b_{1:h-1})\in\Omega_i^{(\mathrm{I}),K}$ and all $(x_{i,h},b_{1:h'})\in\Omega_i^{(\mathrm{II}),K}$. So the assumptions in Lemma A.3 are satisfied. By Lemma A.3, with probability at least $1-p/8$, for all $x_h\in\mathcal{X}_i$, we have

$$\begin{aligned}
&\max_{\varphi}\sum_{t=1}^{T}\prod_{k=1}^{h-1}\pi_i^t(b_k|x_k)\Big\langle\pi_i^t(\cdot|x_h)-\varphi\diamond\pi_i^t(\cdot|x_h),\widetilde{L}_{(x_h,b_{1:h-1})}^t\Big\rangle.\\
\le\ &\frac{2A_i\log(8X_iA_i/p)}{\eta w_{b_{1:h-1}}}+\eta\sum_{t=1}^{T}M_{b_{1:h-1}}^t\overline{L}\Big\langle\pi_i^t(\cdot|x_k),\widetilde{L}_{(x_h,b_{1:h-1})}^t(\cdot)\Big\rangle\\
=\ &\frac{2A_i\log(8X_iA_i/p)}{\eta\prod_{k\in\overline{W}\cup\{h\}}\pi_{i,k}^{\star,h}(a_k|x_k)}+H\eta\prod_{k\in[h-1]}\pi_{i,k}^t(b_k|x_k)\\
&\times\sum_{t=1}^{T}\Big\langle\pi_i^t(\cdot|x_k),\frac{\mathbf{1}\Big\{(x_h,\cdot)=(x_h^{t,(h,\overline{W})},a_h^{t,(h,\overline{W})})\Big\}\Big(H-h+1-\sum_{h'=h}^{H}r_{i,h'}^{t,(h,\overline{W})}\Big)}{\prod_{k\in\overline{W}\cup\{h\}}\pi_{i,k}^{\star,h}(a_k|x_k)\prod_{k\in[h-1]\setminus\overline{W}}\pi_{i,k}^t(b_k|x_k)}\Big\rangle\\
\le\ &\frac{2A_i\log(8X_iA_i/p)}{\eta\prod_{k\in\overline{W}\cup\{h\}}\pi_{i,k}^{\star,h}(a_k|x_k)}+H^2\eta\sum_{t=1}^{T}\Big\langle\pi_i^t(\cdot|x_k),\frac{\prod_{k\in\overline{W}}\pi_{i,k}^t(b_k|x_k)\mathbf{1}\Big\{(x_h,\cdot)=(x_h^{t,(h,\overline{W})},a_h^{t,(h,\overline{W})})\Big\}}{\prod_{k\in\overline{W}\cup\{h\}}\pi_{i,k}^{\star,h}(a_k|x_k)}\Big\rangle.
\end{aligned}$$

Above, we used (i) our choices of $M^t_{b_{1:h-1}}$ and $w_{b_{1:h-1}}$; (ii) $(|\mathcal{B}^s|+|\mathcal{B}^e|)|\Psi^s| \le A_i^{H+A_i} \le X_i A_i^{A_i}$ and (iii) taking union bound over all infosets. Plugging this into $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}}$, we have

$$
\begin{aligned}
&\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} \\
&= \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\quad \times \max_{\varphi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_h) - \varphi\diamond\pi_i^t(\cdot|x_h), \widetilde{L}^t_{(x_h,b_{1:h-1})} \Big\rangle. \\
&\le \frac{2A_i\log(8X_iA_i/p)}{\eta} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \\
&\quad + H^2\eta \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \\
&\quad \times \sum_{t=1}^{T} \underbrace{\prod_{k\in\overline{W}} \pi_{i,k}^t(b_k|x_k) \Big\langle \pi_i^t(\cdot|x_k), \mathbf{1}\Big\{(x_h,\cdot) = (x_h^{t,(h,\overline{W})}, a_h^{t,(h,\overline{W})})\Big\} \Big\rangle}_{:=\overline{\Delta}_t^{x_h,b_{1:h-1}}}.
\end{aligned}
$$

Letting

$$
\begin{aligned}
\mathrm{I}_h &:= \frac{2A_i\log(8X_iA_i/p)}{\eta} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)}; \\
\mathrm{II}_h &:= H^2\eta \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \sum_{t=1}^{T} \overline{\Delta}_t^{x_h,b_{1:h-1}}.
\end{aligned}
$$

Using Lemma B.4, we have

$$
\begin{aligned}
&\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \\
&= \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k}) \prod_{k\in[h-1]\setminus\overline{W}} \pi_{i,k}^{\star,h}(a_k|x_k)}{\prod_{k\in[h]} \pi_{i,k}^{\star,h}(a_k|x_k)} \\
&\overset{(i)}{=} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{a_h} \frac{\sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k}) \prod_{k\in[h-1]\setminus\overline{W}} \pi_{i,k}^{\star,h}(b_k|x_k)\cdot\pi_{i,h}^{\star,h}(a_h|x_h)}{\prod_{k\in[h]} \pi_{i,k}^{\star,h}(a_k|x_k)} \\
&\overset{(ii)}{=} \sum_{\overline{W}\subset[h-1], |\overline{W}|=(h-1)\wedge(K-1)} A_i^{|\overline{W}|} \times \sum_{x_h,a_h} \Bigg(\prod_{k\in[h]} \pi_{i,k}^{\star,h}(a_k|x_k)\Bigg)^{-1} \\
&\quad \times \sum_{b_{1:h-1}:\mathsf{fill}(\{k\in[h-1]:a_k\neq b_k\}, K\wedge H-1)=\overline{W}} \prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k}) \prod_{k\in[h-1]\setminus\overline{W}} \pi_{i,k}^{\star,h}(b_k|x_k) \prod_{k\in\overline{W}} \pi_{i,k}^{\mathrm{unif}}(b_k|x_k)\pi_{i,h}^{\star,h}(a_h|x_h) \\
&\overset{(iii)}{\le} \sum_{\overline{W}\subset[h-1], |\overline{W}|=(K-1)\wedge(h-1)} A_i^{|\overline{W}|} X_{i,h} A_i \\
&= X_{i,h} \binom{h-1}{(K-1)\wedge(h-1)} A_i^{K\wedge h}.
\end{aligned}
$$

Here, (i) uses that $\pi_{i,h}^{\star,h}(\cdot|x_h)$ is uniform distribution on $\mathcal{A}_i$ and that for $k\in[h-1]\setminus\overline{W}$, we have $a_k = b_k$; (ii) follows from grouping x_h and $b_{1:h}$ by $|\overline{W}|$, where $\pi_{i,k}^{\mathrm{unif}}$ is the uniform distribution on $\mathcal{A}_i$; (iii) uses Lemma B.4 and the fact that, for each fixed $\overline{W}$ (by relaxing the summation over $b_{1:h-1}$ to the full sum $\sum_{b_{1:h-1}}$) the numerator is no more than the sequence-form of the following policy:

- Sample recommended action b_k from $\pi_{i,k}^{\star,h}(\cdot|x_k)$ if step $k \in \overline{W}$. Otherwise, sample recommended action b_k from $\pi_{i,k}^{\text{unif}}(\cdot|x_k)$.
- "True" actions are sampled from $\phi_k(a_k|x_k, b_{1:k})$ for $k \in [h-1]$. At step h, take action a_h.

Consequently,

$$\sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \le X_{i,h} \binom{h-1}{K \wedge h - 1} A_i^{K \wedge h}. \tag{35}$$

So we have

$$\text{I}_h \le \frac{2A_i \log(8X_iA_i/p)}{\eta} X_{i,h} \binom{h-1}{(K-1) \wedge (h-1)} A_i^{K \wedge h}.$$

For II_h, observe that the random variables $\overline{\Delta}_t^{x_h, b_{1:h-1}}$ satisfy the following:

- $\overline{\Delta}_t^{x_h, b_{1:h-1}} = \prod_{k \in \overline{W}} \pi_{i,k}^t(b_k|x_k) \left\langle \pi_i^t(\cdot|x_k), \mathbf{1}\left\{(x_h, \cdot) = (x_h^{t,(h,\overline{W})}, a_h^{t,(h,\overline{W})})\right\} \right\rangle \in [0,1]$;
- Let $\mathcal{F}_{t-1}$ be the σ-algebra containing all information until π^t is sampled, then

$$\begin{aligned}
&\mathbb{E}\left[\overline{\Delta}_t^{x_h, b_{1:h-1}} | \mathcal{F}_{t-1}\right] \\
&= \prod_{k \in \overline{W}} \pi_{i,k}^t(b_k|x_k) \mathbb{E}\left[\sum_{a \in \mathcal{A}_i} \pi_i^t(a|x_h) \mathbf{1}\left\{(x_h, a) = (x_h^{t,(h,\overline{W})}, a_h^{t,(h,\overline{W})})\right\} \Big| \mathcal{F}_{t-1}\right] \\
&= \prod_{k \in \overline{W}} \pi_{i,k}^t(b_k|x_k) \sum_{a \in \mathcal{A}_i} \pi_i^t(a|x_h) \mathbb{P}^{((\pi_{i,k}^{\star,h})_{k \in \overline{W} \cup \{h\}} (\pi_{i,k}^t)_{k \in [h-1] \setminus \overline{W}}) \times \pi_{-i}^t}(x_h^{t,(h,\overline{W})} = x_h, a_h^{t,(h,\overline{W})} = a) \\
&= \prod_{k \in \overline{W}} \pi_{i,k}^t(b_k|x_k) \sum_{a \in \mathcal{A}_i} \pi_i^t(a|x_h) \left(\prod_{k \in \overline{W}} \pi_{i,k}^{\star,h}(a_k|x_k) \cdot \prod_{k \in [h-1] \setminus \overline{W}} \pi_{i,k}^t(a_k|x_k) \cdot \pi_{i,h}^{\star,h}(a|x_h) p_{1:h}^{\pi_{-i}^t}(x_h) \right) \\
&= \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k) \cdot \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k) \cdot p_{1:h}^{\pi_{-i}^t}(x_h),
\end{aligned}$$

  where the last equation is because $\prod_{k \in \overline{W}} \pi_{i,k}^t(b_k|x_k) \cdot \prod_{k \in [h-1] \setminus \overline{W}} \pi_{i,k}^t(a_k|x_k) = \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k)$;

- The conditional variance $\mathbb{E}[(\overline{\Delta}_t^{x_h, b_{1:h-1}})^2 | \mathcal{F}_{t-1}]$ can be bounded as

$$\begin{aligned}
\mathbb{E}\left[\left(\overline{\Delta}_t^{x_h, b_{1:h-1}}\right)^2 \Big| \mathcal{F}_{t-1}\right] &\le \mathbb{E}\left[\overline{\Delta}_t^{x_h, b_{1:h-1}} \Big| \mathcal{F}_{t-1}\right] \\
&= \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k) \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h),
\end{aligned}$$

  where we used $\overline{\Delta}_t^{x_h, b_{1:h-1}} \in [0,1]$ almost surely.

Therefore, we can apply Freedman's inequality (Lemma A.1) and union bound to get that for any fixed $\lambda \in (0,1]$, with probability at least $1 - p/8$, the following holds simultaneously for all $(h, x_{i,h}, b_{1:h-1})$:

$$\sum_{t=1}^T \overline{\Delta}_t^{x_h, b_{1:h-1}} \le (\lambda + 1) \sum_{t=1}^T \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k) \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h) + \frac{C \log(X_iA_i/p)}{\lambda},$$

where $C > 0$ is some absolute constant. Plugging this bound into II_h yields that,

$$\text{II}_h \le H^2 \eta \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)}$$

$$\times \left[(\lambda+1) \sum_{t=1}^{T} \prod_{k \in [h-1]} \pi_{i,k}^{t}(b_k|x_k) \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k) \cdot p_{1:h}^{\pi_{-i}^{t}}(x_h) + \frac{C \log(X_i A_i/p)}{\lambda} \right].$$

Note that by Corollary D.2, for fixed t and any $\phi \in \Phi_i^K$,

$$\sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \prod_{k \in [h-1]} \pi_{i,k}^{t}(b_k|x_k) \cdot p_{1:h}^{\pi_{-i}^{t}}(x_{i,h}) \le 1.$$

so we have

$$\max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^{T} \prod_{k \in [h-1]} \pi_{i,k}^{t}(b_k|x_k) \cdot p_{1:h}^{\pi_{-i}^{t}}(x_{i,h}) \le T.$$

Moreover, by the previous bound (35), for any $\phi \in \Phi_i^K$,

$$\sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \le X_{i,h} \binom{h-1}{K \wedge h-1} A_i^{K \wedge h}.$$

Using these two inequalities, we can get that

$$\mathrm{II}_h \le H^2 \eta (\lambda+1) T + \frac{C H^2 \eta \log(X_i A_i/p)}{\lambda} X_{i,h} \binom{h-1}{K \wedge h-1} A_i^{K \wedge h}.$$

Taking summation over $h \in [H]$, we have

$$\begin{aligned}
&\sum_{h=1}^{H} \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} = \sum_{h=1}^{H} (\mathrm{I}_h + \mathrm{II}_h) \\
\le\ & H^3 \eta (\lambda+1) T + \left(\frac{2A_i \log(8X_i A_i/p)}{\eta} + \frac{C H^2 \eta \log(X_i A_i/p)}{\lambda} \right) \sum_{h=1}^{H} X_{i,h} \binom{h-1}{(K-1) \wedge (h-1)} A_i^{K \wedge h} \\
\le\ & H^3 \eta (\lambda+1) T + \left(\frac{2A_i \log(8X_i A_i/p)}{\eta} + \frac{C H^2 \eta \log(X_i A_i/p)}{\lambda} \right) X_i \binom{H-1}{K \wedge H-1} A_i^{K \wedge H},
\end{aligned}$$

for all $\lambda \in (0, 1]$. Choosing $\lambda = 1$, we have,

$$\sum_{h=1}^{H} \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} \le 2H^3 \eta T + \left(\frac{2A_i \log(8X_i A_i/p)}{\eta} + C H^2 \eta \log(X_i A_i/p) \right) X_i \binom{H-1}{K \wedge H-1} A_i^{K \wedge H}.$$

Then, choosing $\eta = \sqrt{\binom{H}{K \wedge H} A_i^{K \wedge H+1} X_i \iota / (H^3 T)}$ and using $\binom{H-1}{K \wedge H-1} \le \binom{H}{K \wedge H}$ we have

$$\begin{aligned}
\sum_{h=1}^{H} \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{swap}} \le\ & 2\sqrt{H^3 \binom{H}{K \wedge H} A_i^{K \wedge H+1} X_i T \iota} \\
& + \mathcal{O}\left(\binom{H}{K \wedge H} A_i^{K \wedge H} X_i \iota \sqrt{\frac{H \binom{H}{K \wedge H} A_i^{K \wedge H+1} X_i \iota}{T}} \right)
\end{aligned}$$

with probability at least $1 - p/8$, where $\iota = \log(8X_i A_i/p)$ is a log factor. □

G.4.2 Proof of Lemma G.5: Bound on $\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}}$

Proof. We can rewrite $\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}}$ as

$$\begin{aligned}
&\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} \\
&= \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)}
\end{aligned}$$

$$\times \sum_{t=1}^{T} \underbrace{\prod_{k\in \overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k) \prod_{k=1}^{h-1} \pi^t_i(b_k|x_k) \Big\langle \pi^t_{i,h}(\cdot|x_{i,h}), L^t_{i,h}(x_{i,h},\cdot) - \widetilde{L}^t_{(x_h,b_{1:h-1})} \Big\rangle}_{:=\widetilde{\Delta}^{x_h,b_{1:h-1}}_t}.$$

Here, for fixed $x_h \in \mathcal{X}_i$ and $b_{1:h-1}$, let $W = \{k \in [h-1] : b_k \neq a_k\}$ and $\overline{W} = \mathsf{fill}(W, (h-1) \wedge (K-1))$. Observe that the random variables $\widetilde{\Delta}^{x_h,b_{1:h-1}}_t$ satisfy the following:

- By the definition of $\widetilde{L}$ in Algorithm 6, we can rewrite $\widetilde{\Delta}^{x_h,b_{1:h-1}}_t$ as

$$\widetilde{\Delta}^{x_h,b_{1:h-1}}_t = \prod_{k\in \overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k) \prod_{k=1}^{h-1} \pi^t_i(b_k|x_k) \Big\langle \pi^t_{i,h}(\cdot|x_{i,h}), L^t_{i,h}(x_{i,h},\cdot) \Big\rangle$$
$$- \prod_{k\in \overline{W}} \pi^t_{i,k}(b_k|x_k) \Big\langle \pi^t_{i,h}(\cdot|x_{i,h}), \mathbf{1}\Big\{(x_h,\cdot) = (x^{t,(h,\overline{W})}_h, a^{t,(h,\overline{W})}_h)\Big\} \cdot \Big(H - h + 1 - \sum_{h'=h}^{H} r^{t,(h,\overline{W})}_{i,h'}\Big) \Big\rangle;$$

- $\widetilde{\Delta}^{x_h,b_{1:h-1}}_t \le \Big\langle \pi^t_{i,h}(\cdot|x_{i,h}), L^t_{i,h}(x_{i,h},\cdot) \Big\rangle \le H;$

- $\mathbb{E}[\widetilde{\Delta}^{x_h,b_{1:h-1}}_t|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information until π^t is sampled. This also can be seen from the unbiasedness of $\widetilde{L}$;

- The conditional variance $\mathbb{E}[(\widetilde{\Delta}^{x_h,b_{1:h-1}}_t)^2|\mathcal{F}_{t-1}]$ can be bounded as

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\widetilde{\Delta}^{x_h,b_{1:h-1}}_t\Big)^2\Big|\mathcal{F}_{t-1}\Big] \\
&\le \mathbb{E}\Bigg[\Big(H-h+1-\sum_{h'=h}^{H} r^{t,(h,\overline{W})}_{i,h'}\Big)^2 \Big(\prod_{k\in\overline{W}} \pi^t_{i,k}(b_k|x_k) \Big\langle \pi^t_{i,h}(\cdot|x_{i,h}), \mathbf{1}\Big\{(x_h,\cdot) = (x^{t,(h,\overline{W})}_h, a^{t,(h,\overline{W})}_h)\Big\}\Big\rangle\Big)^2 \Big|\mathcal{F}_{t-1}\Bigg] \\
&\overset{(i)}{\le} H^2 \prod_{k\in\overline{W}} \pi^t_{i,k}(b_k|x_k) \cdot \mathbb{E}\Big[\Big\langle \pi^t_{i,h}(\cdot|x_{i,h}), \mathbf{1}\Big\{(x_h,\cdot) = (x^{t,(h,\overline{W})}_h, a^{t,(h,\overline{W})}_h)\Big\}\Big\rangle \Big|\mathcal{F}_{t-1}\Big] \\
&= H^2 \prod_{k\in\overline{W}} \pi^t_{i,k}(b_k|x_k) \cdot \mathbb{E}\Big[\sum_{a\in\mathcal{A}_i} \pi^t_i(a|x_k) \mathbf{1}\Big\{(x_h,a) = (x^{t,(h,\overline{W})}_h, a^{t,(h,\overline{W})}_h)\Big\} \Big|\mathcal{F}_{t-1}\Big] \\
&= H^2 \prod_{k\in\overline{W}} \pi^t_{i,k}(b_k|x_k) \cdot \sum_{a\in\mathcal{A}_i} \pi^t_i(a|x_k) \mathbb{P}^{((\pi^{\star,h}_{i,k})_{k\in\overline{W}\cup\{h\}}(\pi^t_{i,k})_{k\in[h-1]\setminus\overline{W}})\times\pi^t_{-i}}(x^{t,(h,\overline{W})}_h = x_h, a^{t,(h,\overline{W})}_h = a) \\
&= H^2 \prod_{k\in\overline{W}} \pi^t_{i,k}(b_k|x_k) \cdot \sum_{a\in\mathcal{A}_i} \pi^t_i(a|x_k) \Big(\prod_{k\in\overline{W}} \pi^{\star,h}_{i,k}(a_k|x_k) \cdot \prod_{k\in[h-1]\setminus\overline{W}} \pi^t_{i,k}(a_k|x_k) \cdot \pi^{\star,h}_{i,h}(a|x_h) p^{\pi^t_{-i}}_{1:h}(x_h)\Big) \\
&\overset{(ii)}{\le} H^2 \prod_{k\in[h-1]} \pi^t_{i,k}(b_k|x_k) p^{\pi^t_{-i}}_{1:h}(x_{i,h}) \prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k).
\end{aligned}$$

  Here, (i) uses $\Big\langle \pi^t_{i,h}(\cdot|x_{i,h}), \mathbf{1}\Big\{(x_h,\cdot) = (x^{t,(h,\overline{W})}_h, a^{t,(h,\overline{W})}_h)\Big\}\Big\rangle \in [0,1]$; (ii) is because $\prod_{k\in\overline{W}} \pi^t_{i,k}(b_k|x_k) \cdot \prod_{k\in[h-1]\setminus\overline{W}} \pi^t_{i,k}(a_k|x_k) = \prod_{k\in[h-1]} \pi^t_{i,k}(b_k|x_k)$.

Therefore, we can apply Freedman's inequality and union bound to get that for any fixed $\lambda \in (0, 1/H]$, with probability at least $1 - p/8$, the following holds simultaneously for all $(h, x_{i,h}, b_{1:h-1})$:

$$\sum_{t=1}^{T} \widetilde{\Delta}^{x_h,b_{1:h-1}}_t \le \lambda H^2 \prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k) \sum_{t=1}^{T} \prod_{k\in[h-1]} \pi^t_{i,k}(b_k|x_k) p^{\pi^t_{-i}}_{1:h}(x_{i,h}) + \frac{C\log(X_iA_i/p)}{\lambda},$$

where $C > 0$ is some absolute constant. Plugging this bound into $\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}}$ yields that, for all $h \in [H]$,

$$\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} \le \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)}$$

$$\times \left[\lambda H^2 \prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k) \sum_{t=1}^{T} \prod_{k\in[h-1]} \pi_{i,k}^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h}) + \frac{C\log(X_iA_i/p)}{\lambda}\right]$$

$$\le \lambda H^2 \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^{T} \prod_{k\in[h-1]} \pi_{i,k}^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h})$$

$$+ \frac{C\log(X_iA_i/p)}{\lambda} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)}.$$

By Corollary D.2, for fixed t and any $\phi \in \Phi_i^K$,

$$\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \prod_{k\in[h-1]} \pi_{i,k}^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \le 1.$$

So we have

$$\lambda H^2 \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^{T} \prod_{k\in[h-1]} \pi_{i,k}^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \le \lambda H^2 T.$$

Moreover, by the inequality (35) in the proof of Lemma G.4, we have for any $\phi \in \Phi_i^K$,

$$\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \le X_{i,h} \binom{h-1}{(K-1)\wedge(h-1)} A_i^{K\wedge h}.$$

Plugging these bounds into $\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}}$ yields that, with probability at least $1-p/8$, for all $h \in [H]$,

$$\mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} \le \lambda H^2 T + \frac{C\log(X_iA_i/p)}{\lambda} X_{i,h} \binom{h-1}{(K-1)\wedge(h-1)} A_i^{K\wedge h}.$$

Taking summation over $h \in [H]$, we have

$$\begin{aligned}
\sum_{h=1}^{H} \mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} &\le \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda} \sum_{h=1}^{H} X_{i,h} \binom{h-1}{(K-1)\wedge(h-1)} A_i^{K\wedge h} \\
&\le \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda} X_i \binom{H-1}{K\wedge H-1} A_i^{K\wedge H} \\
&\le \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda} X_i \binom{H}{K\wedge H} A_i^{K\wedge H},
\end{aligned}$$

for all $\lambda \in (0, 1/H]$. Choose $\lambda = \min\left\{\frac{1}{H}, \sqrt{\frac{CX_i\binom{H}{K\wedge H}A_i^{K\wedge H}\log(X_iA_i/p)}{H^3T}}\right\}$, we obtain the bound

$$\sum_{t=1}^{T} \mathrm{BIAS}_{1,h}^{T,\mathrm{swap}} \le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H}X_iT\iota} + H\binom{H}{K\wedge H}A_i^{K\wedge H}X_i\iota\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor. □

G.4.3 Proof of Lemma G.6: Bound on $\mathrm{BIAS}_{2,h}^{T,\mathrm{swap}}$

Proof. For fixed x_h and $b_{1:h-1}$, let $W = \{k \in [h-1] : b_k \neq a_k\}$ and $\overline{W} = \mathsf{fill}(W, (K-1)\wedge(h-1))$ We can rewrite $\mathrm{BIAS}_{2,h}^{T,\mathrm{swap}}$ as

$$\mathrm{BIAS}_{2,h}^{T,\mathrm{swap}} = \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})$$

$$
\begin{aligned}
&\times \max_{\varphi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \Big\langle \varphi \diamond \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h, b_{1:h-1})} - L^t_{i,h}(x_{i,h}, \cdot) \Big\rangle \\
&= \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\quad \times \max_{\varphi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \sum_{b_h} \pi_{i,h}^t(b_h|x_h) \Big(\widetilde{L}^t_{(x_h, b_{1:h-1})}(\varphi(b_h)) - L^t_{i,h}(x_{i,h}, \varphi(b_h)) \Big) \\
&\le \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_i} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \\
&\quad \times \max_{\phi_h'} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \sum_{b_h, a_h} \pi_{i,h}^t(b_h|x_h) \phi_h'(a_h|x_h, b_{1:h}) \Big(\widetilde{L}^t_{(x_h, b_{1:h-1})}(a_h) - L^t_{i,h}(x_{i,h}, a_h) \Big) \\
&\overset{(i)}{=} \max_{\phi \in \Phi_i^K} \sum_{x_{i,h}, a_h} \sum_{b_{1:h-1}} \sum_{b_h} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi_k(a_k|x_k, b_{1:k}) \\
&\quad \times \sum_{t=1}^{T} \prod_{k=1}^{h} \pi_i^t(b_k|x_k) \Big(\widetilde{L}^t_{(x_h, b_{1:h-1})}(a_h) - L^t_{i,h}(x_{i,h}, a_h) \Big) \\
&= \max_{\phi \in \Phi_i^K} \sum_{x_{i,h}, a_h} \sum_{b_{1:h-1}} \sum_{b_h} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)} \\
&\quad \times \underbrace{\sum_{t=1}^{T} \prod_{k=1}^{h} \pi_i^t(b_k|x_k) \Big(\widetilde{L}^t_{(x_h, b_{1:h-1})}(a_h) - L^t_{i,h}(x_{i,h}, a_h) \Big) \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k)}_{:= \widetilde{\Delta}_t^{x_h, b_{1:h}, a_h}}.
\end{aligned}
$$

Here, (i) comes from the fact that the inner max over ϕ_h' and the outer max over $\phi_{1:h-1}$ are separable and thus can be merged into a single max over $\phi_{1:h}$.

Observe that the random variables $\widetilde{\Delta}_t^{x_{i,h}, b_{1:h}, a_h}$ satisfy the following:

- By the definition of $\widetilde{L}$ in Algorithm 6, we can rewrite $\widetilde{\Delta}_t^{x_h, b_{1:h}, a_h}$ as

$$
\begin{aligned}
&\widetilde{\Delta}_t^{x_h, b_{1:h}, a_h} \\
&= \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^t(b_k|x_k) \mathbf{1}\left\{ (x_h, a_h) = (x_h^{t,(h,\overline{W})}, a_h^{t,(h,\overline{W})}) \right\} \cdot \left(H - h + 1 - \sum_{h'=h}^{H} r_{i,h'}^{t,(h,\overline{W})} \right) \\
&\quad - \prod_{k \in [h]} \pi_{i,k}^t(b_k|x_k) \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^{\star,h}(a_k|x_k) L^t_{i,h}(x_{i,h}, a_h).
\end{aligned}
$$

- $\widetilde{\Delta}_t^{x_h, b_{1:h}, a_h} \in [-H, H]$.
- $\mathbb{E}[\widetilde{\Delta}_t^{x_h, b_{1:h}, a_h} | \mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information until π^t is sampled.
- The conditional variance $\mathbb{E}[(\widetilde{\Delta}_t^{x_h, b_{1:h}, a_h})^2 | \mathcal{F}_{t-1}]$ can be bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\Delta}_t^{x_h, b_{1:h-1}, a_h} \right)^2 \Big| \mathcal{F}_{t-1} \right] \\
&\le \mathbb{E}\left[\left(H - h + 1 - \sum_{h'=h}^{H} r_{i,h'}^{t,(h,\overline{W})} \right)^2 \left(\prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^t(b_k|x_k) \mathbf{1}\left\{ (x_h, a_h) = (x_h^{t,(h,\overline{W})}, a_h^{t,(h,\overline{W})}) \right\} \right)^2 \Big| \mathcal{F}_{t-1} \right] \\
&\le H^2 \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^t(b_k|x_k) \mathbb{P}^{((\pi_{i,k}^{\star,h})_{k \in \overline{W} \cup \{h\}} (\pi_{i,k}^t)_{k \in [h-1] \setminus \overline{W}}) \times \pi_{-i}^t} \left((x_h^{t,(h,\overline{W})}, a_h^{t,(h,\overline{W})}) = (x_h, a_h) \right) \\
&= H^2 \prod_{k \in \overline{W} \cup \{h\}} \pi_{i,k}^t(b_k|x_k) \prod_{k \in \overline{W}} \pi_{i,k}^{\star,h}(a_k|x_k) \cdot \prod_{k \in [h-1] \setminus \overline{W}} \pi_{i,k}^t(a_k|x_k) \cdot \pi_{i,h}^{\star,h}(a_h|x_h) p_{1:h}^{\pi_{-i}^t}(x_{i,h})
\end{aligned}
$$

$$\overset{(i)}{=} H^2 \prod_{k\in[h]} \pi^t_{i,k}(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) \prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k).$$

Here, (i) is because $\prod_{k\in\overline{W}\cup\{h\}} \pi^t_{i,k}(b_k|x_k) \cdot \prod_{k\in[h-1]\setminus\overline{W}} \pi^t_{i,k}(a_k|x_k) = \prod_{k\in[h]} \pi^t_{i,k}(b_k|x_k)$.

Therefore, we can apply Freedman's inequality and union bound to get that for any fixed $\lambda \in (0, 1/H]$, with probability at least $1 - p/8$, the following holds simultaneously for all $(h, x_{i,h}, b_{1:h}, a_h)$:

$$\sum_{t=1}^T \widetilde{\Delta}_t^{x_h, b_{1:h}, a_h} \le \lambda H^2 \prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k) \sum_{t=1}^T \prod_{k\in[h]} \pi^t_{i,k}(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) + \frac{C\log(X_iA_i/p)}{\lambda},$$

where $C > 0$ is some absolute constant. Plugging this bound into $\mathrm{BIAS}^{T,\mathrm{swap}}_{2,h}$ yields that, for all $h \in [H]$ and $\phi \in \Phi_i^K$,

$$\begin{aligned}
\mathrm{BIAS}^{T,\mathrm{swap}}_{2,h} \le& \max_{\phi\in\Phi_i^K} \sum_{x_h,a_h} \sum_{b_{1:h}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^h \phi_k(a_k|x_k, b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k)} \\
&\times \left[\lambda H^2 \prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k) \sum_{t=1}^T \prod_{k\in[h]} \pi^t_{i,k}(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) + \frac{C\log(X_iA_i/p)}{\lambda}\right] \\
\le& \max_{\phi\in\Phi_i^K} \lambda H^2 \sum_{x_h,a_h} \sum_{b_{1:h}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^h \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^T \prod_{k\in[h]} \pi^t_{i,k}(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) \\
&+ \max_{\phi\in\Phi_i^K} \frac{C\log(X_iA_i/p)}{\lambda} \sum_{x_h,a_h} \sum_{b_{1:h}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^h \phi_k(a_k|x_k, b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k)}.
\end{aligned}$$

By Corollary D.2, for fixed t and any $\phi \in \Phi_i^K$,

$$\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \prod_{k\in[h-1]} \pi^t_{i,k}(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) \le 1.$$

so we have

$$\begin{aligned}
&\sum_{x_h,a_h} \sum_{b_{1:h}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^h \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^T \prod_{k\in[h]} \pi^t_{i,k}(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) \\
=& \sum_{x_h} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \sum_{t=1}^T \prod_{k\in[h-1]} \pi^t_{i,k}(b_k|x_k) p_{1:h}^{\pi^t_{-i}}(x_{i,h}) \le T.
\end{aligned}$$

Moreover, by the inequality (35) in the proof of Lemma G.4, we have

$$\begin{aligned}
&\sum_{x_h,a_h} \sum_{b_{1:h}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^h \phi_k(a_k|x_k, b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k)} \\
\overset{(i)}{=}& A_i \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\le K-1}(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})}{\prod_{k\in\overline{W}\cup\{h\}} \pi^{\star,h}_{i,k}(a_k|x_k)} \\
\le& X_{i,h} \binom{h-1}{K\wedge H-1} A_i^{K\wedge h+1}.
\end{aligned}$$

Above, (i) sums over a_h and b_h. Plugging these bounds into $\mathrm{BIAS}^{T,\mathrm{swap}}_{2,h}$ yields that, with probability at least $1 - p/8$, for all $h \in [H]$,

$$\mathrm{BIAS}^{T,\mathrm{swap}}_{2,h} \le \lambda H^2 T + \frac{C\log(X_iA_i/p)}{\lambda} X_{i,h} \binom{h-1}{K\wedge H-1} A_i^{K\wedge h+1}.$$

Taking summation over $h \in [H]$, we have

$$\sum_{h=1}^H \mathrm{BIAS}^{T,\mathrm{swap}}_{2,h} \le \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda} \sum_{h=1}^H X_{i,h} \binom{h-1}{(K-1)\wedge(h-1)} A_i^{K\wedge h+1}$$

$$\leq \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda} X_i \binom{H-1}{K\wedge H-1} A_i^{K\wedge H+1}$$
$$\leq \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda} X_i \binom{H}{K\wedge H} A_i^{K\wedge H+1},$$

for all $\lambda \in (0, 1/H]$. Choose $\lambda = \min\left\{\frac{1}{H}, \sqrt{\frac{CX_i\binom{H}{K\wedge H}A_i^{K\wedge H+1}\log(X_iA_i/p)}{H^3T}}\right\}$, we obtain the bound

$$\sum_{t=1}^{T} \mathrm{BIAS}_{2,h}^{T,\mathrm{swap}} \leq \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_iT\iota} + H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor. □

G.5 Proof of Lemma G.3

Proof. Throughout the proof, for x_h and $b_{1:h-1}$, let $W = \{k \in [h-1] : b_k \neq a_k\}$ and h' is the maximal element in W, so we have $h' = \tau_K$. Recall that $G_h^{T,\mathrm{ext}}(x_h;\phi)$ (eq. (22)) is defined as

$$G_h^{T,\mathrm{ext}}(x_h;\phi) := \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1:\tau_K}),x_h}.$$

For each $h \in [H]$ and $(x_h, b_{1:h'}) \in \Omega_i^{(\mathrm{II}),K}$, we have

$$\begin{aligned}
\widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1,h'})} &= \max_a \sum_{t=1}^{T}\prod_{k=1}^{h'} \pi_i^t(b_k|x_k)\Big(\big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot)\big\rangle - L_{i,h}^t(x_{i,h},a)\Big)\\
&\leq \max_a \sum_{t=1}^{T}\prod_{k=1}^{h'} \pi_i^t(b_k|x_k)\Big(\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:h'})}\Big\rangle - \widetilde{L}^t_{(x_h,b_{1:h'})}(a)\Big)\\
&\quad + \sum_{t=1}^{T}\prod_{k=1}^{h'} \pi_i^t(b_k|x_k)\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot) - \widetilde{L}^t_{(x_h,b_{1:h'})}\Big\rangle\\
&\quad + \max_a \sum_{t=1}^{T}\prod_{k=1}^{h'} \pi_i^t(b_k|x_k)\Big(\widetilde{L}^t_{(x_h,b_{1:h'})}(a) - L_{i,h}^t(x_{i,h},a)\Big).
\end{aligned}$$

Substituting this into $\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h;\phi)$ yields that

$$\begin{aligned}
\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h;\phi) &= \max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K}) \widehat{R}^{T,\mathrm{ext}}_{(x_h,b_{1:\tau_K}),x_h}\\
&\leq \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}} + \mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} + \mathrm{BIAS}_{2,h}^{T,\mathrm{ext}},
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}} &:= \max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K})\\
&\quad \times \max_a \sum_{t=1}^{T}\prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big(\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:\tau_K})}\Big\rangle - \widetilde{L}^t_{(x_h,b_{1:\tau_K})}(a)\Big),\\
\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} &:= \max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K})\\
&\quad \times \sum_{t=1}^{T}\prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), L_{i,h}^t(x_{i,h},\cdot) - \widetilde{L}^t_{(x_h,b_{1:\tau_K})}\Big\rangle,\\
\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}} &:= \max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K})
\end{aligned}$$

$$\times \max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big(\widetilde{L}^t_{(x_h,b_{1:\tau_K})}(a) - L^t_{i,h}(x_{i,h},a)\Big).$$

The bounds for $\sum_{h=1}^{H} \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}}$, $\sum_{h=1}^{H} \mathrm{BIAS}_{1,h}^{T,\mathrm{ext}}$, and $\sum_{h=1}^{H} \mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}$ are given in the following three Lemmas (proofs deferred to Appendix G.5.1, G.5.2, and G.5.3) respectively.

Lemma G.7 (Bound on $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}}$)**.** *If we choose learning rates as*

$$\eta_{x_h} = \sqrt{\binom{H}{K\wedge H} X_i A_i^{K\wedge H+1} \log(8X_iA_i/p)/(H^3T)}$$

for all $x_h \in \mathcal{X}_i$ (same with (10)). Then with probability at least $1-p/4$, we have

$$\sum_{h=1}^{H} \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}} \le \sqrt{H^3 \binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i T\iota}$$
$$+\mathcal{O}\left(\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota \sqrt{\frac{H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota}{T}}\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

Lemma G.8 (Bound on $\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}}$)**.** *With probability at least $1-p/8$, we have*

$$\sum_{h=1}^{H} \mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} \le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i T\iota} + H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

Lemma G.9 (Bound on $\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}$)**.** *With probability at least $1-p/8$, we have*

$$\sum_{h=1}^{H} \mathrm{BIAS}_{2,h}^{T,\mathrm{ext}} \le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H} A_i^{K\wedge H} X_i T\iota} + H\binom{H}{K\wedge H} A_i^{K\wedge H} X_i\iota\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor.

Combining Lemma G.7, G.8, and G.9, we have with probability at least $1-p/2$ that

$$\sum_{h=1}^{H} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} G_h^{T,\mathrm{ext}}(x_h;\phi) \le \sum_{h=1}^{H} \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}} + \sum_{h=1}^{H} \mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} + \sum_{h=1}^{H} \mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}$$
$$\le \sqrt{H^3\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i T\iota}$$
$$+\mathcal{O}\left(\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota \sqrt{\frac{H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota}{T}}\right)$$
$$+\mathcal{O}\left(H\binom{H}{K\wedge H} A_i^{K\wedge H+1} X_i\iota\right).$$

□

G.5.1 Proof of Lemma G.7: Bound on $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}}$

Proof. Recall that $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}}$ is defined as

$$\max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge\tau_K})$$

$$\times \max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big(\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:\tau_K})}\Big\rangle - \widetilde{L}^t_{(x_h,b_{1:\tau_K})}(a)\Big).$$

We first apply regret minimization lemma (Lemma A.3) on $\mathcal{R}_{x_h}$ to give an upper bound of

$$\max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big(\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:\tau_K})}\Big\rangle - \widetilde{L}^t_{(x_h,b_{1:\tau_K})}(a)\Big).$$

For $\mathcal{R}_{x_h}$, Algorithm 7 gives that $M^t_{b_{1:h'}} = \prod_{k=1}^{h'} \pi_i^t(b_k|x_k)$ and $w_{b_{1:h'}} = \prod_{k\in W\cup\{h'+1,\dots,h\}} \pi_{i,k}^{\star,h}(a_k|x_k)$ for any $b_{1:h'} \in \Omega_i^{(\mathrm{II}),K}(x_h)$, where $W = \{k \in [h-1] : b_k \neq a_k\}$. Letting $W(h) = W \cup \{h'+1,\dots,h\}$, η be the learning rate of $\mathcal{R}_{x_h}$ and $\overline{L} = H$. The assumptions in Lemma A.3 are verified in Section G.4.1. So by Lemma A.3, with probability at least $1 - p/8$, we have for all $x_h \in cX_i$,

$$\begin{aligned}
&\max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big(\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:\tau_K})}\Big\rangle - \widetilde{L}^t_{(x_h,b_{1:\tau_K})}(a)\Big)\\
\le\ & \frac{2\log(8X_iA_i/p)}{\eta w_{b_{1:h'}}} + \eta \sum_{t=1}^{T} M^t_{b_{1:h'}} \overline{L} \Big\langle \pi_i^t(\cdot|x_k), \widetilde{L}^t_{(x_h,b_{1:'})}(\cdot)\Big\rangle\\
=\ & \frac{2\log(8X_iA_i/p)}{\eta \prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)} + H\eta \prod_{k=1}^{h'} \pi_i^t(b_k|x_k)\\
&\times \sum_{t=1}^{T} \Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \frac{\mathbf{1}\left\{(x_h,\cdot) = (x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\} \sum_{h''=h}^{H} \left(1 - r_{i,h''}^{t,(h,h',W)}\right)}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k) \prod_{k\in[h']\setminus W} \pi_{i,k}^t(b_k|x_k)}\Big\rangle\\
\le\ & \frac{2\log(8X_iA_i/p)}{\eta \prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)} + H^2\eta \sum_{t=1}^{T} \Big\langle \pi_i^t(\cdot|x_k), \frac{\prod_{k\in W} \pi_{i,k}^t(b_k|x_k) \mathbf{1}\left\{(x_h,a) = (x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\}}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)}\Big\rangle.
\end{aligned}$$

Here, we use (i) our choices of $M^t_{b_{1:h'}}$ and $w_{b_{1:h'}}$; (ii) $(|\mathcal{B}^s| + |\mathcal{B}^e|)|\Psi^e| \le A_i^{H+1} \le X_iA_i$ and (iii) taking union bound over all infosets. Plugging this into $\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}}$, we have

$$\begin{aligned}
&\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}}\\
=\ & \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'})\\
&\times \max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_K} \pi_i^t(b_k|x_k)\Big(\Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \widetilde{L}^t_{(x_h,b_{1:\tau_K})}\Big\rangle - \widetilde{L}^t_{(x_h,b_{1:\tau_K})}(a)\Big)\\
\le\ & \frac{2\log(8X_iA_i/p)}{\eta} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'})}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)}\\
&+ H^2\eta \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'})}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)}\\
&\times \sum_{t=1}^{T} \underbrace{\prod_{k\in W} \pi_{i,k}^t(b_k|x_k) \Big\langle \pi_{i,h}^t(\cdot|x_{i,h}), \mathbf{1}\left\{(x_h,\cdot) = (x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\}\Big\rangle}_{:=\overline{\Delta}_t^{(x_h,b_{1:h'})}}.
\end{aligned}$$

Letting

$$\mathrm{I}_h := \frac{2\log(8X_iA_i/p)}{\eta} \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'})}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)};$$

$$\mathrm{II}_h := H^2\eta \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'})}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)} \sum_{t=1}^{T} \overline{\Delta}_t^{(x_h,b_{1:h'})}.$$

Using Lemma B.4, we have

$$
\begin{aligned}
&\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi^{\star,h}_{i,k}(a_k|x_k)}\\
=&\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})\prod_{k\in[h-1]\setminus W}\pi^{\star,h}_{i,k}(a_k|x_k)}{\prod_{k\in[h]}\pi^{\star,h}_{i,k}(a_k|x_k)}\\
\overset{(i)}{=}&\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{a_h}\frac{\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})\prod_{k\in[h-1]\setminus W}\pi^{\star,h}_{i,k}(b_k|x_k)\cdot\pi^{\star,h}_{i,h}(a_h|x_h)}{\prod_{k\in[h]}\pi^{\star,h}_{i,k}(a_k|x_k)}\\
\overset{(ii)}{=}&\sum_{W\subset[h-1],|W|=K}A_i^{|W|}\sum_{x_h,a_h}\left(\prod_{k\in[h]}\pi^{\star,h}_{i,k}(a_k|x_k)\right)^{-1}\\
&\times\sum_{b_{1:h-1}:\{k\in[h-1]:a_k\neq b_k\}=W}\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})\prod_{k\in[h-1]\setminus W}\pi^{\star,h}_{i,k}(b_k|x_k)\prod_{k\in W}\pi^{\rm unif}_{i,k}(b_k|x_k)\pi^{\star,h}_{i,h}(a_h|x_h)\\
\overset{(iii)}{\le}&\sum_{W\subset[h-1],|W|=K}A_i^{|W|}X_{i,h}A_i\\
=&X_{i,h}\binom{h-1}{K}A_i^{K\wedge H+1}.
\end{aligned}
$$

Here, (i) uses that $\pi^{\star,h}_{i,h}(\cdot|x_h)$ is uniform distribution on $\mathcal{A}_i$ and that for $k\in[h-1]\setminus W$, we have $a_k=b_k$; (ii) follows from grouping x_h and $b_{1:h}$ by $|W|$, where $\pi^{\rm unif}_{i,k}$ is the uniform distribution on $\mathcal{A}_i$; (iii) uses Lemma B.4 and the fact that for any fixed W, (by relaxing the summation over $b_{1:h-1}$ to the full sum $\sum_{b_{1:h-1}}$) the numerator is no more than the sequence-form of the following policy:

- Sample recommended action b_k from $\pi^{\star,h}_{i,k}(\cdot|x_k)$ if step $k\in W$. Otherwise, sample recommended action b_k from $\pi^{\rm unif}_{i,k}(\cdot|x_k)$.
- "True" actions are sampled from $\phi_k(a_k|x_k,b_{1:k})$ for $k\in[h-1]$. At step h, "True" action is sampled from a_h.

Consequently,

$$
\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi^{\star,h}_{i,k}(a_k|x_k)}\le X_{i,h}\binom{h-1}{K}A_i^{K\wedge H+1}. \tag{36}
$$

So we have

$$
\mathrm{I}_h\le\frac{2\log(8X_iA_i/p)}{\eta}X_{i,h}\binom{h-1}{K}A_i^{K\wedge H+1}.
$$

To give an upper bound of I_h, obvserve that the random variables $\overline{\Delta}_t^{(x_h,b_{1:h'})}$ satisfy the following:

- $\overline{\Delta}_t^{(x_h,b_{1:h'})}=\prod_{k\in W}\pi^t_{i,k}(b_k|x_k)\left\langle\pi^t_{i,h}(\cdot|x_{i,h}),\mathbf{1}\left\{(x_h,\cdot)=(x_h^{t,(h,h',W)},a_h^{t,(h,h',W)})\right\}\right\rangle\in[0,1]$.
- Let $\mathcal{F}_{t-1}$ be the σ-algebra containing all information until π^t is sampled, then

$$
\begin{aligned}
&\mathbb{E}\left[\overline{\Delta}_t^{(x_h,b_{1:h'})}|\mathcal{F}_{t-1}\right]\\
=&\prod_{k\in W}\pi^t_{i,k}(b_k|x_k)\mathbb{E}\left[\sum_a\pi^t_{i,h}(a|x_{i,h})\mathbf{1}\left\{(x_h,a)=(x_h^{t,(h,h',W)},a_h^{t,(h,h',W)})\right\}|\mathcal{F}_{t-1}\right]\\
=&\prod_{k\in W}\pi^t_{i,k}(b_k|x_k)\sum_a\pi^t_{i,h}(a|x_{i,h})\mathbb{P}^{((\pi^{\star,h}_{i,k})_{k\in W(h)}(\pi^t_{i,k})_{k\in[h']\setminus W})\times\pi^t_{-i}}(x_h^{t,(h,h',W)}=x_h,a_h^{t,(h,h',W)}=a)
\end{aligned}
$$

$$= \prod_{k\in W} \pi_{i,k}^t(b_k|x_k) \sum_a \pi_{i,h}^t(a|x_{i,h}) \left(\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k) \cdot \prod_{k\in[h']\setminus W} \pi_{i,k}^t(a_k|x_k) \cdot \pi_{i,h}^{\star,h}(a|x_h) p_{1:h}^{\pi_{-i}^t}(x_h) \right)$$

$$= \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k) \prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h),$$

where the last equation is because $\prod_{k\in W} \pi_{i,k}^t(b_k|x_k) \cdot \prod_{k\in[h']\setminus W} \pi_{i,k}^t(a_k|x_k) = \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k)$;

- The conditional variance $\mathbb{E}[(\overline{\Delta}_t^{(x_h,b_{1:h'})})^2|\mathcal{F}_{t-1}]$ can be bounded as

$$\mathbb{E}\left[\left(\overline{\Delta}_t^{(x_h,b_{1:h'})}\right)^2 \middle| \mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\overline{\Delta}_t^{(x_h,b_{1:h'})} \middle| \mathcal{F}_{t-1}\right]$$
$$= \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k) \prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h).$$

 Here, the inequality comes from that $\overline{\Delta}_t^{(x_h,b_{1:h'})} \in [0,1]$.

Therefore, we can apply Freedman's inequality and union bound to get that for any fixed $\lambda \in (0,1]$, with probability at least $1-p/8$, the following holds simultaneously for all $(h, x_{i,h}, b_{1:h-1})$:

$$\sum_{t=1}^T \overline{\Delta}_t^{(x_h,b_{1:h'})} \le (\lambda+1) \sum_{t=1}^T \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k) \prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h) + \frac{C\log(X_iA_i/p)}{\lambda},$$

where $C>0$ is some absolute constant. Plugging this bound into II_h yields that,

$$\mathrm{II}_h \le H^2\eta \max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'})}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)}$$
$$\times \left[(\lambda+1) \sum_{t=1}^T \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k) \prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_h) + \frac{C\log(X_iA_i/p)}{\lambda} \right].$$

Note that for fixed t and any $\phi \in \Phi_i^K$, by Corollary D.2,

$$\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \le 1.$$

So we have

$$\max_{\phi\in\Phi_i^K} \sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \sum_{t=1}^T \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k) p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \le T.$$

Moreover, by the previous bound (36), for any $\phi \in \Phi_i^K$,

$$\sum_{x_h\in\mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\wedge h'})}{\prod_{k\in W(h)} \pi_{i,k}^{\star,h}(a_k|x_k)} \le X_{i,h} \binom{h-1}{K} A_i^{K\wedge H+1}.$$

Using these two inequalities, we can get that

$$\mathrm{II}_h \le H^2\eta(\lambda+1)T + \frac{CH^2\eta\log(X_iA_i/p)}{\lambda} X_{i,h} \binom{h-1}{K} A_i^{K\wedge H+1}.$$

 Taking summation over $h\in[H]$, we have

$$\sum_{h=1}^H \widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}} = \sum_{h=1}^H (\mathrm{I}_h + \mathrm{II}_h)$$
$$\le H^3\eta(\lambda+1)T + \left(\frac{2\log(8X_iA_i/p)}{\eta} + \frac{CH^2\eta\log(X_iA_i/p)}{\lambda} \right) \sum_{h=1}^H X_{i,h} \binom{h-1}{K} A_i^{K\wedge H+1}$$

$$\le H^3\eta(\lambda+1)T + \left(\frac{2\log(8X_iA_i/p)}{\eta} + \frac{CH^2\eta\log(X_iA_i/p)}{\lambda}\right)X_i\binom{H-1}{K\wedge H}A_i^{K\wedge H+1}$$

$$\le H^3\eta(\lambda+1)T + \left(\frac{2\log(8X_iA_i/p)}{\eta} + \frac{CH^2\eta\log(X_iA_i/p)}{\lambda}\right)X_i\binom{H}{K\wedge H}A_i^{K\wedge H+1},$$

for all $\lambda \in (0,1]$. Choosing $\lambda = 1$, we have,

$$\sum_{h=1}^{H}\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}} \le 2H^3\eta T + \left(\frac{2\log(8X_iA_i/p)}{\eta} + CH^2\eta\log(X_iA_i/p)\right)X_i\binom{H-1}{K\wedge H}A_i^{K\wedge H+1}.$$

Then, choosing $\eta = \sqrt{\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota/(H^3T)}$, we have

$$\sum_{h=1}^{H}\widetilde{\mathrm{REGRET}}_h^{T,\mathrm{ext}} \le \sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_iT\iota} + \mathcal{O}\left(\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\sqrt{\frac{H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota}{T}}\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor. □

G.5.2 Proof of Lemma G.8: Bound on $\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}}$

Proof. For fixed x_h and $b_{1:h-1}$, let $W = \{k \in [h-1] : b_k \neq a_k\}$ and h' is the maximal element in W, so we have $h' = \tau_K$. We define $W(h)$ as the set $W \cup \{h'+1, \dots, h\}$. We can rewrite $\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}}$ as

$$\begin{aligned}
&\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} \\
=\ &\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)} \\
&\times\sum_{t=1}^{T}\underbrace{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)\prod_{k=1}^{h'}\pi_{i,k}^t(b_k|x_k)\left\langle\pi_{i,h}^t(\cdot|x_{i,h}),L_{i,h}^t(x_{i,h},\cdot)-\widetilde{L}_{(x_h,b_{1:h'})}^t\right\rangle}_{:=\widetilde{\Delta}_t^{(x_h,b_{1:h'})}}.
\end{aligned}$$

Observe that the random variable $\widetilde{\Delta}_t^{(x_h,b_{1:h'})}$ satisfy the following:

- By the definition of $\widetilde{L}$, we can rewrite $\widetilde{\Delta}_t^{(x_h,b_{1:h'})}$ as

$$\begin{aligned}
\widetilde{\Delta}_t^{(x_h,b_{1:h'})} &= \prod_{k=1}^{h'}\pi_{i,k}^t(b_k|x_k)\left\langle\pi_{i,h}^t(\cdot|x_h),\ L_{(x_h,b_{1:h'})}^t\right\rangle\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k) \\
&- \prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\left\langle\pi_{i,h}^t(\cdot|x_h),\mathbf{1}\left\{(x_h,\cdot)=(x_h^{t,(h,h',W)},a_h^{t,(h,h',W)})\right\}\cdot\left(H-h+1-\sum_{h''=h}^{H}r_{i,h''}^{t,(h,h',W)}\right)\right\rangle;
\end{aligned}$$

- $\widetilde{\Delta}_t^{(x_h,b_{1:h'})} \le \left\langle\pi_i^t(\cdot|x_h),\ L_{(x_h,b_{1:h'})}^t\right\rangle \le H$;

- $\mathbb{E}[\widetilde{\Delta}_t^{(x_h,b_{1:h'})}|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information until π^t is sampled. This also can be seen from the unbiasedness of $\widetilde{L}$;

- The conditional variance $\mathbb{E}[(\widetilde{\Delta}_t^{(x_h,b_{1:h'})})^2|\mathcal{F}_{t-1}]$ can be bounded as

$$\mathbb{E}\left[\left(\widetilde{\Delta}_t^{(x_h,b_{1:h'})}\right)^2\middle|\mathcal{F}_{t-1}\right]$$

$$\le \mathbb{E}\left[\left(H-h+1-\sum_{h''=h}^{H} r_{i,h''}^{t,(h,h',W)}\right)^2 \left(\left\langle \pi_{i,h}^t(\cdot|x_h), \prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\mathbf{1}\left\{(x_h,\cdot)=(x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\}\right\rangle\right)^2 \middle| \mathcal{F}_{t-1}\right]$$

$$\overset{(i)}{\le} H^2 \prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\mathbb{E}\Big[\Big\langle \pi_{i,h}^t(\cdot|x_h), \mathbf{1}\left\{(x_h,\cdot)=(x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\}\Big\rangle \Big| \mathcal{F}_{t-1}\Big]$$

$$= H^2 \prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\mathbb{E}\left[\sum_{a\in\mathcal{A}_i}\pi_{i,h}^t(a|x_h), \mathbf{1}\left\{(x_h,a)=(x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\}\Big\rangle \Big| \mathcal{F}_{t-1}\right]$$

$$= H^2 \prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\cdot\sum_{a\in\mathcal{A}_i}\pi_{i,h}^t(a|x_h)\mathbb{P}^{((\pi_{i,k}^{\star,h})_{k\in W(h)}(\pi_{i,k}^t)_{k\in[h']\setminus W})\times\pi_{-i}^t}\left((x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})=(x_h,a)\right)$$

$$= H^2 \prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\cdot\sum_{a\in\mathcal{A}_i}\pi_{i,h}^t(a|x_h)\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)\cdot\prod_{k\in[h']\setminus W}\pi_{i,k}^t(a_k|x_k)\cdot p_{1:h}^{\pi_{-i}^t}(x_{i,h})$$

$$\overset{(ii)}{\le} H^2 \prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)\cdot\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)\cdot p_{1:h}^{\pi_{-i}^t}(x_{i,h}).$$

Here, (i) uses $\left\langle \pi_{i,h}^t(\cdot|x_h), \mathbf{1}\left\{(x_h,\cdot)=(x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\}\right\rangle\in[0,1]$; (ii) is because

$$\prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\prod_{k\in[h']\setminus W}\pi_{i,k}^t(a_k|x_k)=\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k).$$

Therefore, we can apply Freedman's inequality and union bound to get that for any fixed $\lambda\in(0,1/H]$, with probability at least $1-p/8$, the following holds simultaneously for all $(h, x_{i,h}, b_{1:h-1})$:

$$\sum_{t=1}^{T}\widetilde{\Delta}_t^{(x_h,b_{1:h'})}\le\lambda H^2\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})+\frac{C\log(X_iA_i/p)}{\lambda},$$

where $C>0$ is some absolute constant. Plugging this bound into $\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}}$ yields that, for all $h\in[H]$,

$$\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}}\le\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}$$
$$\times\left[\lambda H^2\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})+\frac{C\log(X_iA_i/p)}{\lambda}\right]$$
$$\le\lambda H^2\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})$$
$$+\frac{C\log(X_iA_i/p)}{\lambda}\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}.$$

Note that for fixed t and any $\phi\in\Phi_i^K$, by Corollary D.2,

$$\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})\le 1.$$

So we have

$$\lambda H^2\max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})\le\lambda H^2T.$$

Moreover, by the inequality (36) in the proof of Lemma G.7, we have for any $\phi\in\Phi_i^K$,

$$\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}\le X_{i,h}\binom{h-1}{K}A_i^{K\wedge H+1}.$$

Plugging these bounds into $\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}}$ yields that, with probability at least $1-p/8$, for all $h \in [H]$,

$$\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} \le \lambda H^2 T + \frac{C\log(X_iA_i/p)}{\lambda} X_{i,h}\binom{h-1}{K} A_i^{K\wedge H+1}.$$

Taking summation over $h \in [H]$, we have

$$\begin{aligned}\sum_{h=1}^{H}\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} \le& \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda}\sum_{h=1}^{H} X_{i,h}\binom{h-1}{K} A_i^{K\wedge H+1} \\ \le& \lambda H^3 T + \frac{C\log(X_iA_i/p)}{\lambda} X_i\binom{H}{K\wedge H} A_i^{K\wedge H+1},\end{aligned}$$

for all $\lambda \in (0, 1/H]$. Choose $\lambda = \min\left\{\frac{1}{H}, \sqrt{\frac{CX_{i,h}\binom{H}{K\wedge H}A_i^{K\wedge H+1}\log(X_iA_i/p)}{H^3T}}\right\}$, we obtain

$$\sum_{t=1}^{T}\mathrm{BIAS}_{1,h}^{T,\mathrm{ext}} \le \mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_iT\iota} + H\binom{H}{K\wedge H}A_i^{K\wedge H+1}X_i\iota\right),$$

where $\iota = \log(8X_iA_i/p)$ is a log factor. □

G.5.3 Proof of Lemma G.9: Bound on $\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}$

Proof. For fixed x_h and $b_{1:h-1}$, let $W = \{k \in [h-1] : b_k \neq a_k\}$ and h' is the maximal element in W, so we have $h' = \tau_K$. We define $W(h)$ as the set $W \cup \{h'+1, \dots, h\}$. We can rewrite $\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}$ as

$$\begin{aligned}\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}} =& \max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1}, b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k, b_{1:k\wedge h'}) \\ &\times \max_a \sum_{t=1}^{T}\prod_{k=1}^{h'}\pi_i^t(b_k|x_k)\Big(\widetilde{L}^t_{(x_h,b_{1:\tau_K})}(a) - L^t_{i,h}(x_{i,h},a)\Big) \\ \le& \max_{\phi\in\Phi_i^K}\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1}, b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k, b_{1:k\wedge h'}) \\ &\times \max_{\phi'_h}\sum_{a_h}\phi'_h(a_h|x_h, b_{1:k\wedge h'})\sum_{t=1}^{T}\prod_{k=1}^{h'}\pi_i^t(b_k|x_k)\Big(\widetilde{L}^t_{(x_h,b_{1:h'})}(a_h) - L^t_{(x_h,b_{1:h'})}(a_h)\Big) \\ =& \max_{\phi\in\Phi_i^K}\sum_{x_{i,h},a_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1}, b_{1:h-1})\prod_{k=1}^{h}\phi_k(a_k|x_k, b_{1:k\wedge h'}) \\ &\times \sum_{t=1}^{T}\prod_{k=1}^{h'}\pi_i^t(b_k|x_k)\Big(\widetilde{L}^t_{(x_h,b_{1:h'})}(a_h) - L^t_{(x_h,b_{1:h'})}(a_h)\Big) \\ =& \max_{\phi\in\Phi_i^K}\sum_{x_{i,h},a_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1}, b_{1:h-1})\frac{\prod_{k=1}^{h}\phi_k(a_k|x_k, b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)} \\ &\times \underbrace{\sum_{t=1}^{T}\prod_{k=1}^{h'}\pi_i^t(b_k|x_k)\Big(\widetilde{L}^t_{(x_h,b_{1:h'})}(a_h) - L^t_{(x_h,b_{1:h'})}(a_h)\Big)\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}_{:=\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h}}\end{aligned}$$

Here, (i) comes from the fact that the inner max over ϕ'_h and the outer max over $\phi_{1:h-1}$ are separable and thus can be merged into a single max over $\phi_{1:h}$.

Observe that the random variable $\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h}$ satisfy the following:

- By the definition of $\widetilde{L}$, we can rewrite $\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h}$ as

$$\begin{aligned}
&\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h} \\
&= \prod_{k\in W} \pi_{i,k}^t(b_k|x_k)\mathbf{1}\left\{(x_h,a_h) = (x_h^{t,(h,h',W)}, a_h^{t,(h,h',W)})\right\}\cdot\left(H-h+1-\sum_{h''=h}^{H} r_{i,h''}^{t,(h,h',W)}\right) \\
&\quad - \prod_{k\in[h']} \pi_{i,k}^t(b_k|x_k)L_{(x_h,b_{1:h'})}^t(a_h)\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k).
\end{aligned}$$

- $\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h} \le H.$
- $\mathbb{E}[\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h}|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information until π^t is sampled.
- The conditional variance $\mathbb{E}[(\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h})^2|\mathcal{F}_{t-1}]$ can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h}\right)^2\Big|\mathcal{F}_{t-1}\right] \\
&\le \mathbb{E}\left[\left(H-h+1-\sum_{h''=h}^{H} r_{i,h''}^{t,(h,h',W)}\right)^2\left(\prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\mathbf{1}\left\{(x_h,a_h)=(x_h^{t,(h,h',W)},a_h^{t,(h,h',W)})\right\}\right)^2\Big|\mathcal{F}_{t-1}\right] \\
&\le H^2\prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\mathbb{P}^{((\pi_{i,k}^{\star,h})_{k\in W(h)}(\pi_{i,k}^t)_{k\in[h']\setminus W})\times\pi_{-i}^t}\left((x_h^{t,(h,h',W)},a_h^{t,(h,h',W)})=(x_h,a_h)\right) \\
&= H^2\prod_{k\in W}\pi_{i,k}^t(b_k|x_k)\prod_{k\in W}\pi_{i,k}^{\star,h}(a_k|x_k)\cdot\prod_{k\in[h']\setminus W}\pi_{i,k}^t(a_k|x_k)\cdot\pi_{i,h}^{\star,h}(a_h|x_h)p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \\
&\overset{(i)}{=} H^2\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k).
\end{aligned}$$

Here, (i) is because $\prod_{k\in[h']\setminus W}\pi_{i,k}^t(a_k|x_k) = \prod_{k\in[h']\setminus W}\pi_{i,k}^t(b_k|x_k)$.

Therefore, we can apply Freedman's inequality and union bound to get that for any fixed $\lambda\in(0,1/H]$, with probability at least $1-p/8$, the following holds simultaneously for all $(h,x_{i,h},b_{1:h},a_h)$:

$$\sum_{t=1}^{T}\widetilde{\Delta}_t^{x_h,b_{1:h'},a_h} \le \lambda H^2\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})+\frac{C\log(X_iA_i/p)}{\lambda},$$

where $C>0$ is some absolute constant. Plugging this bound into $\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}$ yields that, for all $h\in[H]$ and $\phi\in\Phi_i^K$,

$$\begin{aligned}
\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}} &\le \max_{\phi\in\Phi_i^K}\sum_{x_h,a_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}\cdot \\
&\quad\times\left[\lambda H^2\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})+\frac{C\log(X_iA_i/p)}{\lambda}\right] \\
&\le \max_{\phi\in\Phi_i^K}\lambda H^2\sum_{x_h,a_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h}\phi_k(a_k|x_k,b_{1:k\wedge h'})\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h}) \\
&\quad+\max_{\phi\in\Phi_i^K}\frac{C\log(X_iA_i/p)}{\lambda}\sum_{x_h,a_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}.
\end{aligned}$$

Note that for fixed t and any $\phi\in\Phi_i^K$, by Corollary D.2,

$$\sum_{x_h\in\mathcal{X}_{i,h}}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})\le 1.$$

So we have

$$\sum_{x_h,a_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\prod_{k=1}^{h}\phi_k(a_k|x_k,b_{1:k\wedge h'})\sum_{t=1}^{T}\prod_{k\in[h']}\pi_{i,k}^t(b_k|x_k)p_{1:h}^{\pi_{-i}^t}(x_{i,h})\le T.$$

Moreover, by the inequality (36) in the proof of Lemma G.7, we have for any $\phi\in\Phi_i^K$,

$$\begin{aligned}&\sum_{x_h,a_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}\\&=\sum_{x_h}\sum_{b_{1:h-1}}\delta^K(a_{1:h-1},b_{1:h-1})\frac{\prod_{k=1}^{h-1}\phi_k(a_k|x_k,b_{1:k\wedge h'})}{\prod_{k\in W(h)}\pi_{i,k}^{\star,h}(a_k|x_k)}\\&\le X_{i,h}\binom{h-1}{K}A_i^{K\wedge H}\end{aligned}$$

Plugging these bounds into $\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}$ yields that, with probability at least $1-p/8$, for all $h\in[H]$,

$$\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}\le\lambda H^2T+\frac{C\log(X_iA_i/p)}{\lambda}X_{i,h}\binom{h-1}{K}A_i^{K\wedge H}.$$

Taking summation over $h\in[H]$, we have

$$\begin{aligned}\sum_{h=1}^{H}\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}\le&\lambda H^3T+\frac{C\log(X_iA_i/p)}{\lambda}\sum_{h=1}^{H}X_{i,h}\binom{h-1}{K}A_i^{K\wedge H}\\\le&\lambda H^3T+\frac{C\log(X_iA_i/p)}{\lambda}X_i\binom{H}{K}A_i^{K\wedge H},\end{aligned}$$

for all $\lambda\in(0,1/H]$. Choose $\lambda=\min\left\{\frac{1}{H},\sqrt{\frac{CX_i\binom{H}{K\wedge H}A_i^{K\wedge H}\log(X_iA_i/p)}{H^3T}}\right\}$, we obtain the bound

$$\sum_{t=1}^{T}\mathrm{BIAS}_{2,h}^{T,\mathrm{ext}}\le\mathcal{O}\left(\sqrt{H^3\binom{H}{K\wedge H}A_i^{K\wedge H}X_iT\iota}+H\binom{H}{K\wedge H}A_i^{K\wedge H}X_i\iota\right),$$

where $\iota=\log(8X_iA_i/p)$ is a log factor. □

H Additional discussions

H.1 Implementation of K-EFCE by the mediator

Given a K-EFCE, the mediator can implement it as follows. Before the game starts, the mediator samples a product policy from the K-EFCE (which is a correlated policy), and initializes a "deviation counter" for each player at 0. Then, at each round, the mediator by default recommends the sampled actions to all players. After players take their actual action, the mediator increments each player's "deviation counter" by 1 if their action is different from the recommendation. The mediator stops recommending to any player as soon as their counter reaches K. We remark that such an implementation (viewed from the mediator's side) corresponds exactly to the definition of a K-EFCE strategy modification (for the player's side) in Definition 1 and Algorithm 1.

H.2 Requirement on knowing the tree structure in bandit-feedback setting

Under bandit-feedback, in Algorithm 3 and 6, the input balanced exploration policies $\{\pi_i^{\star,h}\}_{h\in[H]}$ depend on the number of children $\mathcal{C}_h(x_{i,h},a_{i,h})$ for all infoset $x_{i,h}$ and action $a_{i,h}$. This requires knowing the structure of each player's game tree (treeplex). A similar requirement is also needed in the Balanced OMD and Balanced CFR algorithm of Bai et al. [5]. We remark that this requirement is relatively mild as the tree structure can be extracted efficiently from just one tree traversal for each player.

H.3 Timeability condition

Our formulation of tree-structure, perfect-recall POMGs is able to express any IIEFG with perfect recall and the additional *timeability* condition [25], a mild condition which roughly speaking requires that infosets for all players combinedly could be partitioned into ordered "layers". Therefore, our results hold for all timeable IIEFGs with perfect recall.

Furthermore, our algorithm K-EFR and Balanced K-EFR does not depend on the joint game tree for all players. Instead, the i^{th} player's version of our algorithms only depends on the i^{th} player's own game tree (which is timeable for any perfect-recall IIEFG). Therefore, our algorithms and theoretical guarantees (when formulated in general IIEFGs with perfect recall) can be generalized directly to any perfect-recall IIEFGs that is not necessarily timeable, similar as existing CFR/OMD type algorithms for external regret minimization [45, 19, 5].

H.4 Comparison between K-EFR and [11, 20] for learning EFCE under full feedback

Celli et al. [11] and its extended version [20] design the first uncoupled no-regret algorithm for computing EFCEs under full feedback. Their algorithms are based on a two-level regret decomposition, which first decomposes the EFCE regret into trigger regrets [22], one for each subtree policy at each infoset, and minimizing each trigger regret via Counterfactual Regret Minimization (CFR). By contrast, our K-EFR utilizes a slightly different decomposition, which decomposes the K-EFCE regret directly into wide-range regrets (7) at each infoset, and uses wide-range regret minimization with WRHEDGE to learn the K-EFCE. We remark that, for learning the EFCE, our approach works by learning the 1-EFCE, which is equivalent to the (trigger definition of the) EFCE in terms of the exact equilibria they define, but a slightly stricter version in terms of ε-approximate equilibria (cf. Proposition C.1).

Morrill et al. [34] considers various forms of correlated equilibria and also uses wide-range regret minimization to learn these equilibria. When specializing in EFCE, we improve its result by a more refined analysis and our new wide-range regret minimization algorithm (WRHEDGE).