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

Ziang Song  
Stanford University  
ziang@stanford.edu

Song Mei  
UC Berkeley  
songmei@berkeley.edu

Yu Bai  
Salesforce Research  
yu.bai@salesforce.com

Abstract

Imperfect-Information Extensive-Form Games (IIIFGs) 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 IIIFGs. 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 $\varepsilon$-approximate $K$-EFCE within $\tilde{O}(\max_i X_i A_i^K / \varepsilon^2)$ 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 $\varepsilon$-approximate $K$-EFCE within $\tilde{O}(\max_i X_i A_i^{K+1} / \varepsilon^2)$ 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 (IIIFGs) [29]. IIIFGs 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 IIIFG, 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 IIIFGs, 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 IIIFGs. 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],

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 \( \varepsilon \)-approximate \( K \)-EFCE within \( O(\max_i X_i A_i^{K+H}/\varepsilon^2) \) iterations under full feedback (Section 4). At \( K = 1 \), our linear in \( \max_i X_i \) dependence improves over the best known \( O(\max_i X_i^2) \) dependence for computing \( \varepsilon \)-approximate EFCE. At \( K > 1 \), this gives a sharp result for efficiently computing (the stricter) \( K \)-EFCE, improving over the best known \( O(\max_i X_i^2 A_i^{K+H}/\varepsilon^2) \) 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 \( \varepsilon \)-approximate \( K \)-EFCE within \( O(\max_i X_i A_i^{K+H+1}/\varepsilon^2) \) 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 SWR\textsc{edge}, 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 \( \varepsilon \)-approximate correlated equilibrium in both [20, 34] scales quadratically in \( \max_{i \in [m]} X_i \). Our \( K \)-EFR algorithm for the
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 (EFCE) [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 Equilibrium (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. Dudik 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, . . . , i}. For a finite set A, we let Δ(A) denote the probability simplex over 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 ∧ n to denote min \{m, n\}.

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

\[
\text{POMG}(m, S, \{X_i\}_{i \in [m]}, \{A_i\}_{i \in [m]}, H, p_0, \{p_h\}_{h \in [H]}, \{r_{i,h}\}_{i \in [m], h \in [H]}),
\]

where S is the state space of size \( |S| = S \); \( X_i \) is the space of information sets (henceforth infosets) for the i\textsuperscript{th} player, which is a partition of \( S \) (i.e., \( x_i \subseteq S \) for all \( x_i \in X_i \), and \( S = \sqcup x_i \in X_i x_i \) where \( \sqcup \) stands for disjoint union) with size \( |X_i| = X_i \), and we also use \( x_i : S \rightarrow X_i \) to denote the i\textsuperscript{th} player’s emission (observation) function; \( A_i \) is the action space for the i\textsuperscript{th} player with size \( |A_i| = A_i \), and we let \( A = A_1 \times \cdots \times A_m \) denote the space of joint actions \( a = (a_1, \ldots, a_m) \); \( H \in \mathbb{Z}_{\geq 1} \) is the time horizon; \( p_0 \in \Delta(S) \) is the distribution of the initial state \( s_1 ; p_h : S \times A \rightarrow \Delta(S) \) are transition probabilities where \( p_h(s_{h+1}|s_h, a_h) \) is the probability of transiting to the next state \( s_{h+1} \) from state-action \( (s_h, a_h) \in S \times A \); and \( r_{i,h} : S \times A \rightarrow [0, 1] \) is the deterministic\(^1\) reward function for the i\textsuperscript{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\(^2\), following [28, 5]. We assume that the game has a tree structure: for any state \( s \in S \), there is a unique step \( h \) and history \( (s_1, a_1, \ldots, s_{h-1}, a_{h-1}, s_h = s) \) to reach \( s \). Precisely, for any policy of the players, for any realization of the game (i.e., trajectory) \( (s'_k, a'_k)_{k \in [H]} \), conditionally to \( s'_l = s \), it almost surely holds that \( l = h \) and \( (s'_1, \ldots, s'_h) = (s_1, \ldots, s_h) \). We also assume history of past infosets and actions \( (x_{i,1}, a_{i,1}, \ldots, x_{i,h-1}, a_{i,h-1}, x_{i,h} = x_i) \) leading to \( x_i \). This requires that \( X_i \) can be partitioned to \( H \) subsets \( (X_{i,h})_{h \in [H]} \) such that \( x_i \) is reachable only at time step \( h \). We define \( X_{i,h} := |X_{i,h}| \). Similarly, the state set \( S \) can be also partitioned into \( H \) subsets \( (S_h)_{h \in [H]} \). As we mostly focus on the i\textsuperscript{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 $x_i(s)$ be 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 $C_{i,h}(x_i,h,a_{i,h}) := \{x \in \mathcal{X}_{i,h'} : (x_i,h,a_{i,h}) \prec x\}$ and $C_{i,h}(x_i,h) := \{x \in \mathcal{X}_{i,h'} : x_i,h \prec x\} = \bigcup_{a_{i,h} \in \mathcal{A}_i} C_{i,h}(x_i,h,a_{i,h})$ 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 $C(x_i,h,a_{i,h}) := C_{h+1}(x_i,h,a_{i,h})$ and $C(x_i,h) := C_{h+1}(x_i,h)$. We denote the set of immediate children.

**Policies**  We use $\pi_i = \{\pi_{i,h}((|x_i,h|))\}_{h \in [H], x_i \in \mathcal{X}_{i,h}}$ to denote a policy of the $i$th player, where each $\pi_{i,h}((|x_i,h|)) \in \Delta(A_i)$ is the action distribution at infoset $x_i,h$. We say $\pi_i$ is a pure policy if $\pi_{i,h}((|x_i,h|)$ 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 $\pi_{i,h} = \{\pi_j\}_{j \in [m], j \neq i}$ denote policies of all players other than the $i$th player. We call $\pi$ a pure product policy if $\pi_i$ is a pure policy for all $i \in [m]$. Let $\Pi_i$ 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 $\pi$ on $\Pi$ induces a correlated policy, which executes as first sampling a product policy $\pi = \{\pi_i\}_{i \in [m]} \in \Pi$ from probability measure $\pi$ and then playing the product policy $\pi$. We also use $\pi$ to denote this policy. A correlated policy $\pi$ can be viewed as a mediator of the game which samples $\pi \sim \pi$ before the game starts, and privately recommends action sampled from $\pi_i((|x_i|))$ 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 $a \in A$, there exists a unique history $(s_1, a_1, \ldots, s_h = s, a_h = a)$ ending with $(s_h = s, a_h = a)$. Given any product policy $\pi$, the probability of reaching $(s_h, a_h)$ at step $h$ can be decomposed as

$$p^\pi_{s_h}(s_h, a_h) = p_{1:h}(s_h) \prod_{i \in [m]} \pi_{i,1:h}(s_h,a_{i,h}),$$

(1)

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

$$p_{1:h}(s_h) := p_1(s_1) \prod_{h=1}^{h-1} p_{h'}(s_{h'}|s_{h'}, a_{h'})$$

(2)

and

$$\pi_{i,1:h}(s_h,a_{i,h}) := \prod_{h'=1}^{h} \pi_{i,h'}(a_{i,h'}|s_{i,h'})$$

(3)

where $(s_{h'}, a_{h'})_{h' \leq h-1}$ is the unique history of states and actions that leads to $s_h$ by the tree structure: $x_i,h = x(s_h)$ 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 $\pi_{i,h'}(x_{i,h'},a_{i,h'}) := \prod_{h''=h}^{h'} \pi_{i,h''}(a_{i,h''}|x_{i,h''})$ for any $1 \leq h \leq h' \leq H$.

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

$$V_{i}^{\pi^T} = \sum_{h=1}^{H} \sum_{(s_h,a_h) = (a_{i,h})_{i \in [m]} \in S_h \times A} p_{1:h}(s_h) \prod_{j=1}^{m} \pi_{j,1:h}(x_j(s_h), a_j,h) r_{i,h}(s_h,a_{i,h})$$

For any sequence of policies $\{\pi^t\}_{t=1}^T$, we also define the counterfactual loss functions [45]

$$L_{i,h}^T(x_i,h,a_{i,h}) := \ell_{i,h}^T(x_i,h,a_{i,h}) + \sum_{h'=h+1}^{H} \sum_{a_{i,h'} \in \mathcal{A}_i} \pi_{i,(h+1):h'}(x_{i,h'},a_{i,h'}) \ell_{i,h}^T(x_{i,h'},a_{i,h'}).$$

(5)
We now introduce the definition of

We now define a

Algorithm 1 Executing modified policy \( \phi \circ \pi_i \)

**Input:** \( K \)-EFCE strategy modification \( \phi \in \Phi^K \) (0 ≤ \( K \) ≤ \( \infty \)), policy \( \pi_i \in \Pi_i \) for the \( i^{th} \) player.

1. Initialize recommendation history \( b = \emptyset \).
2. for \( h = 1, \ldots, H \) do
3. Receive infoset \( x_{i,h} \in X_{i,h} \).
4. if \( b \in \Omega_i^{(1),K}(x_{i,h}) \) then
5. Observe recommendation \( b_h \sim \pi_{i,h}(\cdot|x_{i,h}) \).
6. Take swapped action \( a_h = \phi(x_{i,h}, b, b_h) \).
7. Update recommendation history \( b \leftarrow (b, b_h) \in \mathcal{A}_b^h \).
8. else
9. // Must have \( b \in \Omega_i^{(1),K}(x_{i,h}) \), do not observe recommendation from \( i \).
10. Take action \( a_h = \phi(x_{i,h}, b) \).

Intuitively, \( L^i_{1,h}(x_{i,h}, a_{i,h}) \) 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 \( \pi_j \) at all time steps, and the \( i^{th} \) player’s own policy \( \pi_i^t \) from step \( h + 1 \) onward. We will omit the \( i \) subscript and use \( L_h \) 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 \( \pi^f = \{ \pi_j \}_{j \in [m]} \) in each iteration and observe the counterfactual loss functions \( \{ L^i_{1,h}(x_{i,h}, a_{i,h}) \}_{i,h,x_{i,h},a_{i,h}} \) 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 \( \mathcal{A}^h := \{(b_1, \ldots, b_h) \mid b_h \in A_i, \forall h' \leq h \} \), we categorize all possible recommendation histories (henceforth rechistories) at each infoset \( x_{i,h} \in X_{i,h} \) (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** (\( \leq K-1 \) deviations happened) at \( x_{i,h} \) is any action sequence \( b_{1:h-1} \in \mathcal{A}^{h-1}_h \) such that \( \sum_{k=1}^{h-1} 1\{a_k \neq b_k\} \leq K-1 \), where \( (a_1, \ldots, a_{h-1}) \) is the unique sequence of actions leading to \( x_{i,h} \). Let \( \Omega_h^{(1),K}(x_{i,h}) \) 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 \( b_{1:h'} \in \mathcal{A}^{h'}_h \) with length \( h' < h \) such that \( \sum_{k=1}^{h'-1} 1\{a_k \neq b_k\} = K - 1 \) and \( a_{h'} \neq b_{h'} \), where \( (a_1, \ldots, a_{h-1}) \) is the unique sequence of actions leading to \( x_{i,h} \). Let \( \Omega_h^{(1),K}(x_{i,h}) \) denote the set of all Type-II rechistories at \( x_{i,h} \).

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

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

\(^4\)This is implementable (and slightly more general than) when the full game (transitions and rewards) is known.
Let $\Phi^K_i$ denote the set of all possible $K$-EFCE strategy modifications for any $0 \leq K \leq \infty$. Formally, for any $\phi \in \Phi^K_i$ and any pure policy $\pi_i \in \Pi_i$, we define the modified policy $\phi \circ \pi_i$ as in Algorithm 1.

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

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

Definition 2 ($K$-EFCE). A correlated policy $\pi$ is an $\varepsilon$-approximate $K$ Extensive-Form Correlated Equilibrium ($K$-EFCE) if

$$K\text{-EFCEGap}(\pi) := \max_{i \in [m]} \max_{\phi \in \Phi^K_i} \left( V_i^{\phi \circ \pi} - V_i^\pi \right) \leq \varepsilon. $$

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

### 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], 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

$$R^K_{t, \pi} := \max_{\phi \in \Phi^K_i} \sum_{s=1}^{T} \left( V_i^{\phi \circ \pi \times \pi_{-i}} - V_i^{\pi_{-i}} \right).$$

(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 $R^K_{t, \pi}$, by maintaining a regret minimizer $R_{x_{i,t}}$ (using algorithm REGAL) at each infoset $x_{i,t} \in \mathcal{X}_i$ that is responsible for outputting the policy $\pi^t_i(\cdot | x_{i,t}) \in \Delta_{A_i}$ (Line 8) which combine to give the overall policy $\pi^t_{i}$ for the $t$-th iteration.

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

$$\max_{\phi \in \Phi^*} \sum_{s=1}^{T} \prod_{k=1}^{s_{b_{1:t}}} \pi^t_i(b_k | x_k) \left( \left( \pi^t_i(\cdot | x_{i,t}) - \phi \circ \pi^t_i(\cdot | x_{i,t}), L^t_{i,h}(x_{i,t}) \right) \right)$$

(7)

for all possible Type-I rechistories $b_{1:t-1} \in \mathcal{O}_i^{(1)}$ simultaneously, where $\prod_{k=1}^{s_{b_{1:t}}} \pi^t_i(b_k | x_k) := S^t_{b_{1:t-1}}$ is the time-selection function (i.e. a weight function) associated with this $b_{1:t-1}$ (cf. Line 5), and $\Psi^* = \{ \psi : A_i \to A_i \}$ is the set of all swap modifications from the action set $A_i$ onto itself. (An analogous regret for Type-II rechistories is also controlled by $R_{x_{i,t}}$.)

Controlling these “local”
Algorithm 2 \( K \)-EFR with full feedback (\( i \)th player’s version)

**Input:** Algorithm REGALG for minimizing wide range regret; learning rates \( \{\eta_{x_i,h}\} \) for each player.

1. Initialize regret minimizers \( \{R_{x_i,h}\} \) for each player with REGALG and learning rate \( \eta_{x_i,h}\).

2. **for** iteration \( t = 1, \ldots, T \) **do**

3. **for** \( h = 1, \ldots, H \) **do**

4. **for** \( x_i,h \in X_i \) **do**

5. Compute \( S^t_{h-1} = \prod_{b_{i,h-1}}^{t-1} \pi_i^t(b_{i,h-1} | x_i) \) for all \( b_{i,h-1} \in \Omega_i^{(h-1)}, K(x_i,h) \).

6. Compute \( S^t_{h,b} = \prod_{k=1}^{t'} \pi_i^t(b_{i,h} | x_i) \) for all \( b_{i,h} \in \Omega_i^{(h)}, K(x_i,h) \).

7. **for** \( i,h \), \( \text{OBSERVE-TIMESELECTION} \)

8. **Set policy** \( \pi_i \) for each player.

9. **Observe counterfactual losses** \( L^T_{i,h}(x_i,h,a) \) (depending on \( \pi_i \) and \( \pi_{-i} \); cf. (5)).

10. **for** all \( x_i,h \in X_i \) **do**

11. **OBSERVE-LOSS** \( \{L^T_{i,h}(x_i,h,a)\} \).

**Output:** Policies \( \{\pi^T_i\} \).

---

regrets at each \( x_i,h \) guarantees that the overall \( K \)-EFR regret is bounded, by the \( K \)-EFR 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

\[
\eta_{x_i,h} = \sqrt{\left(\frac{H}{K \wedge H}\right)X_iA_i^{K \wedge H} \log A_i / (H^2T)}
\]  

for all \( x_i,h \in X_i \). 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 \( K \)-EFR from full feedback). For any \( 0 < K \leq \infty, \varepsilon \in (0, H] \), 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 \( \pi^t = \{\pi_i^t\} \) denote the joint policy of all players at the \( t \)th iteration. Then the average policy \( \pi = \text{Unif}(\{\pi_i^T\}) \) satisfies \( K \)-EFRGap(\( \pi \)) \leq \varepsilon, as long as the number of iterations

\[
T \geq \mathcal{O}\left(\left(\frac{H}{K \wedge H}\right) \max_{x_i \in [m]} X_i A_i^{K \wedge H} T / \varepsilon^2 \right),
\]

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

In the special case of \( K = 1 \), Theorem 3 shows that \( K \)-EFR can compute an \( \varepsilon \)-approximate 1-EFCE within \( \mathcal{O}(\max_{x_i \in [m]} X_i A_i / \varepsilon) \) iterations. This improves over the existing \( \mathcal{O}(\max_{x_i \in [m]} X_i^2 A_i^2 / \varepsilon^2) \) 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 \( \mathcal{O}(\max_{x_i \in [m]} X_i^{4 / 3} A_i^{4 / 3} / \varepsilon^{4 / 3}) \), we achieve lower \( X_i \) dependence (though worse \( \varepsilon \) dependence).

For \( 1 < K \leq \infty \), Theorem 3 gives a sharp \( \mathcal{O}(\max_{x_i \in [m]} X_i A_i^{K \wedge H} / \varepsilon^2) \) iteration complexity for computing \( K \)-EFR. This improves over the \( \mathcal{O}(\max_{x_i \in [m]} X_i^2 A_i^{K \wedge H} / \varepsilon^2) \) rate of EFR [34] instantiated to the \( K \)-EFR problem. Also, note that although the term \( A_i^{K \wedge H} \) is exponential in \( K \) (for \( K \leq H \)), 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 \( K \)-EFR strategy modification. Apart from learning equilibria, Algorithm 2 also achieves a low \( K \)-EFR regret when controlling the \( i \)th player only and facing potentially adversarial opponents:

\[\text{More precisely, Anagnostiades et al. [3, Corollary 4.17] proves an } \mathcal{O}(\max_{x_i \in [m]} \|\pi_i\|^2 / \varepsilon A_i) \text{ iteration complexity, which specializes to the above rate, as for any } \delta > 0 \text{ a game with } \max_{x_i \in [m]} \|\pi_i\|_1 \geq X_i^{-\delta} \text{ can be constructed.}\]
Algorithm 3 Loss estimator for Type-II rechistories via Balanced Sampling (i'th player's version)

**Input:** Policy $\pi_i^t$, $\pi_i^{,h}$. Balanced exploration policies $(\pi_i^{,h})_{h \in [H]}$.

1. for $K \leq h < H$, $W \subseteq [h']$ with $|W| = K$ and ending in $h'$ do
2. Set policy $\pi_i^{t,(h,h'),W} \leftarrow (\pi_i^{,h})_{k \in [h'] \setminus W} \cdot (\pi_i^{t})_{k \in [h'] \cap W} \cdot (\pi_i^{(h,+,h+1) \cdot H}$.
3. Play $\pi_i^{t,(h,h'),W} \times \pi_i^{t}$ for one episode, observe trajectory $x_{i,h}, a_h, x_{i,h}, a_h, \ldots, x_{i,h}, a_h, x_{i,h}$.
4. for all $(x_{i,h+1}, b_{i,h+1}) \in \Omega_i^{(II),K}$ do
5. Find $(x_{i,1}, a_1) \leftarrow \cdots \leftarrow (x_{i,h-1}, a_{h-1}) \leftarrow x_{i,h}$.
6. Set $W \leftarrow \{ k \in [h'] : b_k \neq a_k \}$.
7. Construct loss estimator for all $a \in A_i$

$$\tilde{L}_i^{t}(x_{i,h}, b_{i,h}, a_i) \leftarrow \frac{1}{\pi_i^{t,(h,h',W)}(x_{i,h}, a_i)} \sum_{h'=h}^{H} \left( 1 - r_{i,h',W} \right). \tag{9}$$

**Output:** Loss estimators $\{ \tilde{L}_i^{t}(x_{i,h}, b_{i,h}, a_i) \}_{(x_{i,h}, b_{i,h}, a_i) \in \Omega_i^{(II),K}}$

$R_{t,K}^T \leq \tilde{O}((H/K \wedge H) X_i A_i^{K \wedge H T})$ (Corollary F.1). In particular, the $\tilde{O}(\sqrt{T})$ scaling is optimal up to log factors, due to the fact that $R_{t,K}^T \geq R_{t,0}^T$ (i.e. the vanilla regret) and the known lower bound $R_{t,0}^T \geq \Omega(\sqrt{T})$ in II-EFGs [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 $|C_h(x_{i,h'}, a_{i,h'})|$ 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 $1 \leq h \leq H$, we define $\pi_i^{*,h}$, the (i'th player's) balanced exploration policy for layer $h$ as

$$\pi_i^{*,h}(a_{i,h'}|x_{i,h'}) := \frac{|C_h(x_{i,h'}, a_{i,h'})|}{|C_h(x_{i,h'})|} \text{ for all } (x_{i,h'}, a_{i,h'}) \in X_{i,h'} \times A_i, h' \leq h - 1,$$

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

Note that there are $H$ such policies, one for each layer $h$. We remark that the construction of $\pi_i^{*,h}$ 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 5; 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 \( \{ \pi^e_i, h, \pi^b_i \} \) that are \textit{interlaced concatenations} of the current \( \pi_i \) and the balanced policy \( \pi^b_i \), and play one episode using each policy against \( \pi^e_{i,j} \).

• Line 7: Construct loss estimators \( \{ L_{x_{i,h}, h, b_{i,h'}, a} \} \) by (9), where for each \( x_{i,h} \) and \( b_{1,h'} \in \Omega^{(1)}_{i,K} (x_{i,h}) \) is an unbiased estimator of counterfactual losses \( \{ L^e_{i} (x_{i,h}, a) \} \in A_i \) that will be used by Algorithm 7 to be fed into the regret minimization algorithm REGALG.

We remark that the sampling policies \( \{ \pi^e_i, h, h', W \} \) in Algorithm 3 are \textit{interlaced concatenations} of \( \pi_i \) and \( \pi^b_i \) 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 \textit{time-selection aware sampling}: Each loss estimator \( L^e_{x_{i,h}, b_{1,h', h}^\dagger} \) achieves low variance relative to the corresponding time selection function \( S^e_{i,h', h} \). Further, there is an \textit{efficient sharing} of sampling policies, as here roughly \((H_K^H x_i A_i^{K_H H})\) loss estimators (one for each \( (x_{i,h}, b_{1,h'}) \)) are constructed using only (a much lower number of) \( H(K_K H) \) 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

\[
\eta_{x_{i,h}} = \sqrt{\left( H_{K_K H}^H x_i A_i^{K_K H + 1} \right) \log(8 \sum_{i \in [m]} x_i A_i / p) / (H^3 T)}, \quad T = H. \quad (10)
\]

SWRHEDGE is a non-trivial extension of WRHEDGE to the stochastic setting, as in each round it admits \textit{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 \( L^e_{x_{i,h}, b_{1,h'}} \) for each \( b_{1,h'} \in \Omega^{(1)}_{i,K} (x_{i,h}) \) (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 \( 0 \leq K \leq \infty \), \( \varepsilon \in (0, H] \) 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 \( \pi_i = \{ \pi^e_i \}_{i \in [m]} \) denote the joint policy of all players at the \( t \)th iteration. Then, with probability at least \( 1 - p \), the correlated policy \( \pi = \text{Unif} (\{ \pi^e_i \}_{i=1}^T) \) satisfies \( K \)-EFCEGap(\( \pi \)) \leq \varepsilon \), as long as \( T \geq \mathcal{O}(H^3 H_{K_K H}^H x_i A_i^{K_K H + 1}) / \varepsilon^2 \). The total number of episodes played is

\[
3m H_{H_{K_K H}}^H T = \mathcal{O} \left( \left( m \left( H_{K_K H}^H \right)^2 \max_{i \in [m]} x_i A_i^{K_K H + 1} \right) / \varepsilon^2 \right),
\]

where \( \lambda = \log(8 \sum_{i \in [m]} x_i A_i / p) \) is a log factor and \( \mathcal{O}(\cdot) \) hides \( \mathcal{O}(\cdot) \) factors.

To our best knowledge, Theorem 5 provides the first result for learning \( K \)-EFCE under bandit feedback. The sample complexity \( \mathcal{O}(\left( H_{K_K H}^H \right)^2 \max_{i \in [m]} (x_i A_i^{K_K H + 1}) / \varepsilon^2) \) (ignoring \( m, H \) factors) has only an \( \left( H_{K_K H}^H \right) A_i \) additional factor over the iteration complexity in the full feedback setting (Theorem 3), which is natural—The \( \left( H_{K_K H}^H \right) \) 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 \( \mathcal{O}(\max_{i \in [m]} x_i A_i^{K_H H + 1}) / \varepsilon^2 \). We remark that the linear in \( x_i \) dependence at all \( K \geq 0 \) is optimal, as the sample complexity lower bound for the \( K = 0 \) case (learning NFCCEs) is already \( \Omega(\max_{i \in [m]} x_i A_i / \varepsilon^2) [5] \). Also, the policies \( \{ \pi^e_i \}_{i=1}^T \) 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 \( \pi^e_i, h, h', W \) actually used are not \( \pi^e_i \).

**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 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 \( \pi^*_{h} \) (Lemma B.4) to control the variance of the loss estimators \( \tilde{L}_t^{(x,a,b)} \), 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 \( \infty \)-EFCE and the BCE. S.M. is supported by NSF grant DMS-2210827.

References


Appendix

Table of Contents

A Technical tools 14
   A.1 Technical lemmas .............................................. 14
   A.2 Wide-range regret minimization ............................... 14
   A.3 Stochastic wide-range regret minimization .................. 16

B Properties of the game 18
   B.1 Basic properties ............................................. 18
   B.2 Balanced exploration policy ................................. 20

C Relationship between $K$-EFCE and existing equilibria 21
   C.1 Proof of Proposition C.1 ...................................... 22
   C.2 Proof of Proposition C.2 ...................................... 24
   C.3 Proof of Proposition C.3 ...................................... 25

D Properties of $K$-EFCE strategy modifications 26

E Regret decomposition for $K$-EFCE regret 28

F Proofs for Section 4 32
   F.1 Proof of Lemma F.1 ............................................ 33
   F.2 Proof of Lemma F.2 ............................................ 35

G Proofs for Section 5 37
   G.1 Sample-based loss estimator for Type-I rechistories ........ 37
   G.2 Algorithm description of Balanced $K$-EFR .................. 37
   G.3 Proof of Theorem 5 ............................................. 39
   G.4 Proof of Lemma G.2 ........................................... 41
      G.4.1 Proof of Lemma G.4: Bound on $\text{REGRET}^{T,\text{swap}}_h$ ... 43
      G.4.2 Proof of Lemma G.5: Bound on $\text{BIAS}^{T,\text{swap}}_{1,h}$ .... 46
      G.4.3 Proof of Lemma G.6: Bound on $\text{BIAS}^{T,\text{swap}}_{2,h}$ .... 48
   G.5 Proof of Lemma G.3 ........................................... 51
      G.5.1 Proof of Lemma G.7: Bound on $\text{REGRET}^T_h$ ........... 52
      G.5.2 Proof of Lemma G.8: Bound on $\text{BIAS}^{T,\text{ext}}_{1,h}$ .... 56
      G.5.3 Proof of Lemma G.9: Bound on $\text{BIAS}^{T,\text{ext}}_{2,h}$ .... 58

H Additional discussions 60
   H.1 Implementation of $K$-EFCE by the mediator ................ 60
   H.2 Requirement on knowing the tree structure in bandit-feedback setting .... 60
   H.3 Timeability condition ........................................ 61
   H.4 Comparison between $K$-EFR and [11, 20] for learning EFCE under full feedback 61
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

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

A.2 Wide-range regret minimization

**Time selection functions** For each element \( b \) in a finite set \( B \), 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 \( B \) be the non-intersection union of \( B^\ast \) and \( B^e \), where \( B^\ast \) is called the set of swap time selection indexes and \( B^e \) is called the set of external time selection indexes.

**Strategy modifications** We denote the set of swap modification function set \( \Psi^\ast = \{\psi: [A] \rightarrow [A]\} \) and external modification function set \( \Psi^e = \{\psi: [A] \rightarrow [A]: \exists a \in [A], \text{s.t.} \psi(b) = a, \forall b \in [A]\} \). Given a modification \( \psi: [A] \rightarrow [A] \) and a strategy \( p^t \in \Delta([A]) \), we denote \( \psi(p^t) \in \Delta([A]) \) to be the modified strategy with \( (\psi \circ p^t)(a) = \sum_{b \in \psi^{-1}(a)} p^t(b) \). Given a modification \( \psi: [A] \rightarrow [A] \), 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 protocol and learning goals** We consider the following interaction protocol: at each iteration \( t \), the learner receives the time selection set \( \{S^b_t\}_{b \in B} \subseteq [0, 1] \), 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

\[
\sum_{t=1}^T (\ell^t - \langle \psi \circ p^t, \ell^t \rangle)
\]

for each pair \((b, \psi) \in \mathcal{I} := (B^\ast \times \Psi^\ast) \cup (B^e \times \Psi^e)\).

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 \( B^\ast \) and external index set \( B^e \); Swap strategy modification set \( \Psi^\ast \) and external strategy modification set \( \Psi^e \).

1. Let \( \mathcal{I} \) := \((B^\ast \times \Psi^\ast) \cup (B^e \times \Psi^e)\). Initialize \( S^b_0 \leftarrow 0 \) for all \( b \in B^\ast \cup B^e \), and

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

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

2. for iteration \( t = 1, \ldots, T \) do

3. (OBSERVE_TIME_SELECTION) Receive time selection functions \( \{S^b_t\}_{b \in B^\ast \cup B^e} \).

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

\[
q^t(b, \psi) \propto q^{t-1}(b, \psi) \exp \left\{ \eta \exp(-\eta\|\ell^{t-1}\|_\infty) S^{t-1}_b(p^{t-1}, \ell^{t-1}) - \eta S^{t-1}_b \psi(p^{t-1}, \ell^{t-1}) \right\}.
\]

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

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 \( \{S_b^t\}_{b \in B^r \cup B^s, t \in [T]} \) 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 \( b \in B^r \cup B^s \), \( a \in [A] \) and \( t \in [T] \). Let \( p^t \) be given as in Algorithm 4 with learning rate \( \eta \in (0, \infty) \). Then we have

\[
\sup_{\psi \in \Psi^*} \sum_{t=1}^{T} S_b^t \left( \langle p^t, \ell^t \rangle - \langle \psi \circ p^t, \ell^t \rangle \right) \leq \sum_{t=1}^{T} \eta \|\ell^t\|_\infty S_b^t \left( \langle p^t, \ell^t \rangle \right) + \frac{\log((|B^s| + |B^r|)|\Psi^*|)}{\eta}, \quad \forall b \in B^r,
\]

\[
\sup_{\psi \in \Psi^*} \sum_{t=1}^{T} S_b^t \left( \langle p^t, \ell^t \rangle - \langle \psi \circ p^t, \ell^t \rangle \right) \leq \sum_{t=1}^{T} \eta \|\ell^t\|_\infty S_b^t \left( \langle p^t, \ell^t \rangle \right) + \frac{\log((|B^s| + |B^r|)|\Psi^*|)}{\eta}, \quad \forall b \in B^s.
\]

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

\[
L_b^t := \sum_{t'=1}^{t} S_b^{t'} \left( \langle p^{t'}, \ell^{t'} \rangle \right),
\]

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

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

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

\[
w^t(b, \psi) := (|\Psi^*|\{b \in B^r\} + |\Psi^*|\{b \in B^s\}) \exp \left\{ \eta \sum_{t'=1}^{t} \exp(-\eta\|\ell^{t'}\|_\infty)S_b^{t'} \left( \langle \psi, \ell^{t'} \rangle \right) - \eta L^t(b, \psi) \right\},
\]

and hence \( w^0(b, \psi) \) is given by \( |\Psi^*|\{b \in B^r\} + |\Psi^*|\{b \in B^s\} \). We further let \( W^t := \sum_{(b, \psi) \in I} w^t(b, \psi) \). 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 \( W^t < W^{t-1} \) for all \( t \geq 1 \). In fact, by \( \exp(-\eta x) < 1 - (1 - \exp(-\eta\|\ell^{t'}\|_\infty))x/\|\ell^{t'}\|_\infty \) and \( \exp(\eta x) < 1 + (\exp(\eta\|\ell^{t'}\|_\infty) - 1)x/\|\ell^{t'}\|_\infty \) for any \( \eta \in (0, \infty) \) and \( x \in [0, \|\ell^{t'}\|_\infty] \), we have

\[
W^t = \sum_{(b, \psi) \in I} w^t(b, \psi) = \sum_{(b, \psi) \in I} w^{t-1}(b, \psi) \exp \left\{ \eta S_b^{t-1} \left( \langle \psi, \ell^{t-1} \rangle \right) - \langle \eta \|\ell^{t-1}\|_\infty p^{t-1} - \psi \circ p^{t-1}, \ell^{t-1} \rangle \right\}
\]

\[
\leq \sum_{(b, \psi) \in I} w^{t-1}(b, \psi) \left( 1 - \frac{(1 - \exp(-\eta\|\ell^{t-1}\|_\infty))S_b^{t-1} \left( \langle \psi, \ell^{t-1} \rangle \right)}{\|\ell^{t-1}\|_\infty} \right) + \frac{1 - \exp(-\eta\|\ell^{t-1}\|_\infty)}{\|\ell^{t-1}\|_\infty} W^{t-1} \sum_{(b, \psi) \in I} q^t(b, \psi) S_b^{t} \left( \langle \psi, \ell^{t} \rangle \right)
\]

\[
\leq W^{t-1}.
\]

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

\[
p^{t-1} = p^{t-1} \left( \frac{\sum_{(b, \psi) \in I} S_b^{t-1} q^t(b, \psi) M_{\psi}}{\sum_{(b, \psi) \in I} S_b^{t-1} q^t(b, \psi)} \right)
\]

in line 5 in the algorithm, which gives

\[
\sum_{(b, \psi) \in I} q^t(b, \psi) S_b^{t} p^{t} = \sum_{(b, \psi) \in I} q^t(b, \psi) S_b^{t} \langle \psi, p^{t} \rangle.
\]

This proves that \( W^t < W^{t-1} \) for all \( t \geq 1 \).
We have similar results for \( \Psi^* \in [0,1] \cup B^s \), where \( B^s \) is any selection set. Let

\[
\left\{ \begin{array}{l}
\text{Wide-range regret bound for SWRH} \\
\end{array} \right.
\]

and \( \psi \) is arbitrary functions, rescaled time selection functions, and weighting functions. Let \( 0 < T < \infty \) be a parameter (that will serve as an upper bound of all \( w_b M^t_b \| \ell^t_b \|_\infty \)). Assume that (i) \( M^t_b \geq 0 \), \( w_b > 0 \), and \( \ell^t_b \in [0, T/(w_b M^t_b)]^A \) for all \( b \in B^s \cup B^e \), \( a \in [A] \) and \( t \in [T] \); (ii) \( E \left( \ell^t_b, F_{t-1} \right) = \ell^t_b \) for all \( b \in B^s \cup B^e \). Let \( p^t \) be given as in Algorithm 5 with learning rate \( \eta \in (0, \infty) \) and time selection function \( \{ S^t_b \}_{b \in B^s \cup B^e} = \{ w_b M^t_b \}_{b \in B^s \cup B^e} \). Then with probability at least \( 1 - p \), we have

\[
\sup_{\psi \in \Psi^*} \sum^{T}_{t=1} M^t_b \left( \left\langle p^t, \ell^t_b \right\rangle - \left\langle \psi \circ p^t, \ell^t_b \right\rangle \right) \leq \sum^{T}_{t=1} \eta T M^t_b \left( p^t, \ell^t_b \right) + \frac{\log([B^s + |B^e|] |\Psi^*/p)}{\eta w_b}, \quad \forall b \in B^s,
\]

and

\[
\sup_{\psi \in \Psi^*} \sum^{T}_{t=1} M^t_b \left( \left\langle p^t, \ell^t_b \right\rangle - \left\langle \psi \circ p^t, \ell^t_b \right\rangle \right) \leq \sum^{T}_{t=1} \eta T M^t_b \left( p^t, \ell^t_b \right) + \frac{\log([B^s + |B^e|] |\Psi^*/p)}{\eta w_b}, \quad \forall b \in B^e.
\]
We next show that Algorithm 5 Stochastic Wide Range Regret Minimization with Hedge (SWRHEDGE) for all (b, ψ) ∈ I.

1: Let $\mathcal{I} := (B^* × \Psi^*) ∪ (B^c × \Psi^c)$. Initialize $S^0_b \leftarrow 0$ for all $b ∈ B^* ∪ B^c$, and

$$q^0(b, ψ) \leftarrow \frac{|\Psi^*|1 \{b ∈ B^*\} + |\Psi^c|1 \{b ∈ B^c\}}{\sum_{(b, ψ) ∈ \mathcal{I}} [|\Psi^*|1 \{b ∈ B^*\} + |\Psi^c|1 \{b ∈ B^c\}]}$$

for all $(b, ψ) ∈ \mathcal{I}$.

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

3: (OBSERVE_TIME_SELECTION) Receive time selection functions $\{S^t_b\}_{b ∈ B^* ∪ B^c}$.

4: Update distribution over $(b, ψ) ∈ \mathcal{I}$:

$$q^t((b, ψ)) \propto q^{t-1}((b, ψ)) \exp \{\eta \exp(−\eta \mathcal{L}) S^t_b \langle p^{t-1}, \tilde{e}^{t-1}_b \rangle − \eta S^{t-1}_b \langle \psi ∘ p^{t-1}, \tilde{e}^{t-1}_b \rangle\}.$$  

5: Set $p^t ∈ Δ([A])$ as a solution to the equation $p^T = p^T \left(\frac{\sum_{(b, ψ) ∈ \mathcal{I}} S^t_b q^t((b, ψ)) M_{b, ψ}}{\sum_{(b, ψ) ∈ \mathcal{I}} S^t_b q^t((b, ψ))}\right).$

6: if RECOMMEND is called then

7: Output the vector $p^t$.

8: (OBSERVE_LOSS) Receive loss vectors $\{\ell^t_b\}_{b ∈ B^* ∪ B^c}$ (where $E(\tilde{e}^t_b | F_{t-1}) = \ell^t$ doesn’t depend on $b$).

**Proof.** For $(b, ψ) ∈ \mathcal{I}$, we define the cumulative loss w.r.t. $\{S^t_b\}_{t ≥ 1}$ till time $t$ as

$$L^t(b) := \sum_{t'=1}^t S^t_b \langle p^{t'}, \tilde{e}^t_b \rangle,$$

and define the cumulative loss w.r.t. $\{\{S^t_b\}_{t ≥ 1}, ψ\}$ till time $t$ as

$$L^t(b, ψ) := \sum_{t'=1}^t S^t_b \langle ψ ∘ p^{t'}, \tilde{e}^t_b \rangle.$$  

We further define the weight of $\{\{S^t_b\}_{t ≥ 1}, ψ\}$ at the end of time $t$ as

$$w^t(b, ψ) := (|\Psi^*|1 \{b ∈ B^*\} + |\Psi^c|1 \{b ∈ B^c\}) \exp \{\eta \exp(−\eta \mathcal{L}) L^t(b) − \eta L^t(b, ψ)\},$$

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

We next show that $E[W^{t+1} | F_{t-1}] ≤ W^t$ for all $t ≥ 1$. In fact, by $\exp(−\eta x) ≤ 1 − (1 − \exp(−\eta \mathcal{L}))x/\mathcal{L}$ and $E(x) ≤ 1 + (\exp(\eta \mathcal{L}) − 1)x/\mathcal{L}$ for any $\eta ∈ (0, ∞]$ and $x ∈ [0, \mathcal{L}]$, we have

$$W^t = \sum_{(b, ψ) ∈ \mathcal{I}} w^t(b, ψ) = \sum_{(b, ψ) ∈ \mathcal{I}} w^{t-1}(b, ψ) \exp \{\eta S^t_b \langle \exp(−\eta \mathcal{L}) p^{t} − ψ ∘ p^{t}, \tilde{e}^t_b \rangle\} \leq \sum_{(b, ψ) ∈ \mathcal{I}} w^{t-1}(b, ψ) \left(1 − \frac{1 − \exp(−\eta \mathcal{L}) S^t_b \langle ψ ∘ p^{t}, \tilde{e}^t_b \rangle}{\mathcal{L}}\right) \cdot \left(1 + \frac{(1 − \exp(−\eta \mathcal{L}) S^t_b \langle p^{t}, \tilde{e}^t_b \rangle)}{\mathcal{L}}\right)$$

$\leq W^{t-1} − \frac{1 − \exp(−\eta \mathcal{L})}{\mathcal{L}} W^{t-1} \sum_{(b, ψ) ∈ \mathcal{I}} q^t(b, ψ) S^t_b \langle ψ ∘ p^{t}, \tilde{e}^t_b \rangle$

$\quad + \frac{1 − \exp(−\eta \mathcal{L})}{\mathcal{L}} W^{t-1} \sum_{(b, ψ) ∈ \mathcal{I}} q^t(b, ψ) S^t_b \langle p^{t}, \tilde{e}^t_b \rangle.$
Here, (i) follows from \( q^t(b, \psi) = w^{t-1}(b, \psi)/W^{t-1} \) and \( S_b^T \ell_b^T \leq L \) for any \( b \). Using the above inequality, \( \mathbb{E}\left( \ell_b^T | \mathcal{F}_{t-1} \right) = \ell^t \) for any \( b \), and the fact that \( p^t \) solves

\[
p^T = p^T \left( \frac{\sum_{(b,\psi) \in I} S_b^t q^t(b, \psi) M_{\psi}}{\sum_{(b,\psi) \in I} S_b^t q^t(b, \psi)} \right)
\]

in line 5 in the algorithm, which gives

\[
\sum_{(b,\psi) \in I} q^t(b, \psi) S_b^t p^t = \sum_{(b,\psi) \in I} q^t(b, \psi) S_b^t (\psi \circ p^t),
\]

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

\[
\mathbb{E}[W^t] \leq W^0.
\]

Therefore, by Markov inequality, we have with probability at least 1 − \( p \) that

\[
W^t \leq W^0/p.
\]

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

\[
(\Psi^* | 1 \{ b \in B^s \} + \Psi^* | 1 \{ b \in B^c \}) \exp \left\{ \eta \exp(-\eta L) L^t(b) - \eta L^t(b, \psi) \right\} = w^t(b, \psi) \leq W^t \leq W^0/p \leq \sum_{(b', \psi') \in I} w^0(b', \psi') = |\Psi^*| |\Psi^*| (|B^c| + |B^c|)/p.
\]

As a result,

\[
\exp \left\{ \eta \exp(-\eta L) L^t(b) - \eta L^t(b, \psi) \right\} \leq |\Psi^*| (|B^c| + |B^c|)/p, \quad \forall b \in B^s,
\]

\[
\exp \left\{ \eta \exp(-\eta L) L^t(b) - \eta L^t(b, \psi) \right\} \leq |\Psi^*| (|B^c| + |B^c|)/p, \quad \forall b \in B^c.
\]

Note that we have \( 1 \leq \exp(-\eta L) + \eta L \). So we can get that, for any \( b \in B^s \),

\[
L^t(b) - L^t(b, \psi) \leq \exp(-\eta L) L^t(b) - L^t(b, \psi) + \eta L L^t(b) 
\]

\[
\leq \frac{\log(|\Psi^*( |B^c| + |B^c|)/p)}{\eta} + \eta L L^t(b).
\]

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

\[
\sum_{t=1}^T S_b^t \left( \langle p^t, \ell_b^t \rangle - \langle \psi \circ p^t, \rho_b^t \rangle \right) \leq \sum_{t=1}^T \eta L S_b^t \left( \langle p^t, \ell_b^t \rangle \right) + \frac{\log(|\Psi^*( |B^c| + |B^c|)/p)}{\eta} \mathbb{E} \mathbb{E}[W^t] \leq \sum_{t=1}^T \eta L M_b^t \left( \langle p^t, \ell_b^t \rangle \right) + \frac{\log(|\Psi^*( |B^c| + |B^c|)/p)}{\eta w_b}, \quad \forall b \in B^s.
\]

Because \( S_b^t = M_b^t w_b \), dividing by \( w_b \) gives that

\[
\sum_{t=1}^T M_b^t \left( \langle p^t, \ell_b^t \rangle - \langle \psi \circ p^t, \rho_b^t \rangle \right) \leq \sum_{t=1}^T \eta L M_b^t \left( \langle p^t, \ell_b^t \rangle \right) + \frac{\log(|\Psi^*( |B^c| + |B^c|)/p)}{\eta w_b}, \quad \forall b \in B^s.
\]

We have similar results for \( b \in B^c \). Taking supreme over all \( \psi \in \Psi^c \) or \( \psi \in \Psi^s \) proves Lemma A.3. \( \square \)

## 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 \( \{ \pi_{1, h} \}_{j \neq i} \) as in Eq. (3), we define the marginal reaching probability \( p_{1, h}^\pi(s_h) \) and \( p_{1, h}^\pi(x_{i, h}) \) as follows:

\[
p_{1, h}^\pi(s_h) = p_{1, h}(s_h) \prod_{j \in [m], j \neq i} \pi_j,1:h(s_h, o_{j, h}), \quad (12)
\]

\[
p_{1, h}^\pi(x_{i, h}) = \sum_{s_h \in x_{i, h}} p_{1, h}^\pi(s_h). \quad (13)
\]

The following three results give properties of the marginal reaching probability \( p_{1, h}^\pi(x_{i, h}) \), and the counterfactual loss functions \( L_{i, h}^1 \).

---

18
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
\[ \sum_{(x_h, a_h) \in X_{i,h} \times A_i} \pi_{i,1:h}(x_h, a_h)p_{1:h}^{\pi_i}(x_h) = 1. \]

(b) $0 \leq p_{1:h}^{\pi_i}(x_h) \leq 1$ for all $h \in [H], x_h \in X_{i,h}$.

Proof. For (a), notice that
\[ \pi_{i,1:h}(x_h, a_h)p_{1:h}^{\pi_i}(x_h) = \sum_{s_h \in x_h} p_{1:h}(s_h) \cdot \pi_{i,1:h}(x_h, a_h) \cdot \prod_{j \neq i} \pi_{j,1:h-1}(x_j, a_{j-1}, s_{j-1}) \]
\[ = \sum_{s_h \in x_h} \mathbb{P}_{\pi_i,\pi_{-i}}(\text{visit } (s_h, a_h)) = \mathbb{P}_{\pi_i,\pi_{-i}}(\text{visit } (x_h, a_h)). \]

Summing over all $(x_h, a_h) \in X_{i,h} \times A_i$, the right hand side sums to one, thereby showing (a).

For (b), fix any $x_h \in X_{i,h}$. Clearly $p_{1:h}^{\pi_i}(x_h) \geq 0$. Choose any $a_h \in A_i$, and choose policy $\pi_{i,1:h}^{a_i,a_h} \in \Pi_i$ such that $\pi_{i,1:h}^{a_i,a_h}(x_h, a_h) = 1$ (such $\pi_{i,1:h}^{a,i,a_h}$ exists, for example, by deterministically taking all actions prescribed in infost $x_h$ at all ancestors of $x_h$). For this $\pi_{i,1:h}^{a_i,a_h}$, using (a), we have
\[ p_{1:h}^{\pi_i}(x_h) = \pi_{i,1:h}^{a_i,a_h}(x_h, a_h) \cdot p_{1:h}^{\pi_i}(x_h) \leq \sum_{(x_h', a_h') \in X_{i,h} \times A_i} \pi_{i,1:h}^{a_i,a_h'}(x_h', a_h') \cdot p_{1:h}^{\pi_i}(x_h') = 1. \]

This shows part (b).

Corollary B.1. For any policy $\pi_i \in \Pi_i$ and $h \in [H]$, we have
\[ \sum_{(x_h, a_h) \in X_{i,h} \times A_i} \pi_{i,1:h}(x_h, a_h)\ell_{1:h}^t(x_h, a_h) \leq 1. \]

Proof. Notice by definition
\[ \ell_{1:h}^t(x_h, a_h) = \sum_{s_h \in x_h, (a_h, a_i) \neq (a_h, \pi_i)} p_{1:h}(s_h) \prod_{j \neq i} \pi_{j,1:h}^t(x_j, a_j)(s_j, a_j, s_j) (1 - r_h(s_h, a_h)) \leq p_{1:h}^{\pi_i}(x_h), \]
and the result is implied by Lemma B.1 (b).

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

(a) For any policy $\pi_i \in \Pi_i$, we have
\[ \sum_{(x_h, a_h) \in X_{i,h} \times A_i} \pi_{i,1:h}(x_h, a_h)L_{t,h}^i(x_h, a_h) \leq H - h + 1. \]

(b) For any $(h, x_h, a_h)$, we have
\[ 0 \leq L_{t,h}^i(x_h, a_h) \leq p_{1:h}^{\pi_i}(x_h) \cdot (H - h + 1). \]

Proof. Part (a) follows from the fact that
\[ \sum_{(x_h, a_h) \in X_{i,h} \times A_i} \pi_{i,1:h}(x_h, a_h)L_{t,h}^i(x_h, a_h) = \mathbb{E}_{\pi_i,\pi_{-i}} \left[ \sum_{h' = h}^{H} r_{h'} \right] \leq H - h + 1, \]
where the first equality follows from the definition of the loss functions $\ell_h$ and $L_h$ in (4), (5).

For part (b), the nonnegativity follows clearly by definition. For the upper bound, take any policy $\pi_{i,1:h}^{a_i,a_h} \in \Pi_i$ such that $\pi_{i,1:h}^{a_i,a_h}(x_h, a_h) = 1$. We then have
\[ L_{t,h}^i(x_h, a_h) = \pi_{i,1:h}^{a_i,a_h}(x_h, a_h)L_{t,h}^i(x_h, a_h) = \mathbb{E}_{\pi_i,\pi_{-i}} \left[ \mathbf{1}\{\text{visit } x_h, a_h\} \cdot \sum_{h' = h}^{H} r_{h'} \right]. \]
\[ P_{\pi_1, \pi_2} \text{visit } x_h, a_h \cdot P_{\pi_1, \pi_2} \left( \sum_{h'}=h \right) \text{visit } x_{h'}, a_{h'} \leq P_{\pi_1, \pi_2} (x_h, a_h) P_{\pi_2} (x_h) \cdot (H-h+1) = P_{\pi_1, \pi_2} (x_h) \cdot (H-h+1). \]

This proves the lemma. \(\square\)

**B.2 Balanced exploration policy**

Here we collect properties of the balanced exploration policy \(\pi^*_i\) (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 IIEPGs.

We begin by providing an interpretation of the balanced exploration policy \(\pi^*_i\); its inverse \(1/\pi^*_i\) 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 \(p_{i, k}^*(x_{k+1} | x_k, a_k) = |C_h(x_{k+1})| / |C_h(x_k, a_k)|\) (we use the convention that \(|C_h(x_h)| = 1\)). By this definition, \(p_{i, k}^*(x_{k+1} | x_k, a_k)\) is a probability distribution over \(C_h(x_k, a_k)\) 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

\[ p_{i, 1; h}^*(x_h) = \frac{|C_h(x_1)|}{X_{i, h}} \prod_{k=1}^{h-1} p_{i, k}^*(x_{k+1} | x_k, a_k) = \frac{|C_h(x_1)|}{X_{i, h}} \prod_{k=1}^{h-1} |C_h(x_{k+1})| / |C_h(x_k, a_k)|. \]

**Lemma B.3.** For any \((x_h, a_h) \in X_{i, h} \times A_i\), the sequence form of the transition \(p_{i, 1; h}^*(x_h)\) and the sequence form of balanced exploration policy \(\pi_{i, 1; h}^*(x_h, a_h)\) are related by

\[ p_{i, 1; h}^*(x_h) = \frac{1}{X_{i, h} A_i \cdot \pi_{i, 1; h}^*(x_h, a_h)}. \] (15)

Furthermore, for any \(i\)th player’s policy \(\pi_i \in \Pi_i\) and any \(h \in |H|\), we have

\[ \sum_{(x_h, a_h) \in X_{i, h} \times A_i} \pi_{i, 1; h}^*(x_h, a_h) p_{i, 1; h}^*(x_h) = 1. \] (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

\[ \frac{1}{X_{i, h} A_i \cdot \pi_{i, 1; h}^*(x_h, a_h)} = \frac{1}{X_{i, h} A_i} \prod_{k=1}^{h-1} \frac{|C_h(x_k)|}{|C_h(x_k, a_k)|} \times A_i = \frac{|C_h(x_1)|}{X_{i, h}} \prod_{k=1}^{h-1} \frac{|C_h(x_{k+1})|}{|C_h(x_k, a_k)|} = p_{i, 1; h}^*(x_h), \]

where the second equality used the property that \(|C_h(x_h)| = 1\). This proves Eq. (15). The proof of Eq. (16) is similar to the proof of Lemma B.1 (a). \(\square\)

**Lemma B.4 (Balancing property of \(\pi_{i, 1; h}^*\)).** For any \(i\)th player’s policy \(\pi_i \in \Pi_i\) and any \(h \in |H|\), we have

\[ \sum_{(x_h, a_h) \in X_{i, h} \times A_i} \frac{\pi_{i, 1; h}(x_h, a_h)}{\pi_{i, 1; h}^*(x_h, a_h)} = X_{i, h} A_i. \]

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

Lemma B.4 states that \(\pi_{i, 1; h}^*\) 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 \(\pi_{i, 1; h}^*\) with the uniform policy \(\pi_{i, 1; h}^{\text{init}}(x_h, a_h) = 1/A_i^h\), 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 existing 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 $\varepsilon$-approximate EFCE as any correlated policy $\pi$ such that the following trigger gap is at most $\varepsilon$:

$$\text{TriggerGap}(\pi) := \max_{i \in [m]} \max_{(x, a) \in X_i \times A_i} \max_{\tilde{\pi} \in \Pi_i} \left( E_{\pi \sim \pi^*} V_i^{\text{trig}(\pi; \tilde{\pi}, (x, a)) \times \pi^*} - E_{\pi \sim \pi^*} V_i^\pi \right) \leq \varepsilon. \tag{17}$$

Here the trigger policy $\text{trig}(\pi; \tilde{\pi}, (x, a)) \in \Pi_i$ (with triggering sequence $(x, a)$) 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 $\tilde{\pi} \in \Pi_i$ thereafter.

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

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

In particular, 1-EFCEGap$(\pi) = 0$ if and only if TriggerGap$(\pi) = 0$.

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-EFCEGap$(\pi) \leq \varepsilon$ implies TriggerGap$(\pi) \leq \varepsilon$, 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 $K \geq 1$ and shows that they are strict strengthenings of the 1-EFCE. Note that as we consider games with a finite horizon $H$, we have $K$-EFCEGap$ = H$-EFCEGap for all $K \geq H$ (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 $\pi$, we have

$$0\text{-EFCEGap}(\pi) \leq 1\text{-EFCEGap}(\pi) \leq \cdots \leq K\text{-EFCEGap}(\pi) \leq (K + 1)$$.EFCEGap$(\pi) \leq \cdots \leq \infty$-EFCEGap$(\pi)$. In other words, $K$-EFCE are stricter equilibria as $K$ increases: Any $\varepsilon$-approximate $(K + 1)$-EFCE is also an $\varepsilon$-approximate $K$-EFCE for any $\varepsilon \geq 0$ and $K \geq 0$.

Moreover, the converse bounds do not hold, even if multiplicative factors are allowed: For any $0 \leq K < \infty$, there exists a game with $H = K + 1$ and a correlated policy $\pi$ for which $K$-EFCEGap$(\pi) = 0$ but $(K + 1)$-EFCEGap$(\pi) \geq 1/3 > 0$.

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 $\infty$-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 $\Pi^\text{pure}_i$ 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.
We remark though, despite the equivalence between the strategy modifications of well as the recommendation history. This is exactly equivalent to the A “behavioral deviation” ϕ

Step 1. We first show that TriggerGap(π) ≤ 1-EFCEGap(π) for any correlated policy π. We first claim that, for any trigger policy trig(π, i, (x, a)) where πi is a pure policy (at each infoset x, a chooses action πi(x) deterministically.). We prove the two claims separately.

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 πi 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 trig(π, i, (x, a)) where πi is a pure policy (at each infoset x, a chooses action πi(x) deterministically.). We prove the two claims separately.

At a high level, NFCE has the “strongest” form of strategy modifications, which can observe the entire pure policy πi (i.e. full set of recommendations on every infoset). NFCE has the “weakest” form of strategy modifications, which cannot observe any recommendation at all (so that each ϕ ∈ ΦNFCE is equivalent to a pure policy in Πpure).

Definition C.3 (NFCE and NFCE). An ε-approximate {NFCE, NFCE} of a POMG is a correlated policy π such that

\[ (\text{NFCE}, \text{NFCE})\text{Gap}(π) := \max_{\pi \in \{\text{NFCE, NFCE}\}} \max_{\phi \in \Phi_i} \left( \mathbb{E}_{\pi \sim \pi V_i^{(\phi \pi_i)} - \mathbb{E}_{\pi \sim \pi V_i^\pi} \right) \leq \varepsilon. \]

We say π is an (exact) {NFCE, NFCE} if the above holds with ε = 0.

Proposition C.3 (Relationship between K-EFCE and NFCE, NFCE). For any correlated policy π, we have

(a) ∞-EFCEGap(π) ≤ NFCEGap(π), i.e. NFCE is stricter than ∞-EFCE (NFCEGap(π) ≤ ε implies ∞-EFCEGap(π) ≤ ε).

Further, the converse bound does not hold even if multiplicative factors are allowed: there exists a game with H = 2 and a correlated policy π for which ∞-EFCEGap(π) = 0 but NFCEGap(π) > 1/20.

(b) 0-EFCEGap(π) = NFCEGap(π), i.e. 0-EFCE is equivalent to NFCE.

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 πi 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.
Given this claim, for any correlated policy $\pi$, we have as desired

$$\text{TriggerGap}(\pi) = \max_{i \in [m]} \max_{(x,a) \in X_i \times A_i} \max_{\pi \in \Pi_i} \left( \mathbb{E}_{\pi \sim \pi} V^\text{trig}(\pi, \pi_i, (x,a)) \times \pi - \mathbb{E}_{\pi \sim \pi} V^\pi \right)$$

$$\leq \max_{i \in [m]} \max_{\phi \in \Phi_i^1} \left( \mathbb{E}_{\pi \sim \pi} V^\phi(\phi \pi_i, (x,a)) \times \pi - \mathbb{E}_{\pi \sim \pi} V^\pi \right) = 1\text{-EFCEGap}(\pi).$$

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

**Step 2.** We next show that $1\text{-EFCEGap}(\pi) \leq \max_{i \in [m]} X_i A_i \cdot \text{TriggerGap}(\pi)$ for any correlated policy $\pi \in \Delta(\Pi)$. 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 $\phi \circ \pi_i(x_i) \neq \pi_i(x_i)$, we have the decomposition of identity

$$1 = \sum_{h=1}^H \sum_{x_h \in X_{i,h}} \sum_{a_h \in A_i} 1 \left\{ x_h \text{ visited, } a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h \right\}$$

$$+ \sum_{x_H \in X_{1:H}} \sum_{a_H \in A_i} 1 \left\{ x_H \text{ visited, } a_{1:H} \text{ recomd., and } \phi(x_H, a_{1:H}) = a_H \right\}.$$  

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

$$V^\phi(\phi \pi_i, \pi_{-i}) - V^\pi = (\mathbb{E}_{\phi \pi_i} - \mathbb{E}_{\pi}) \left[ \sum_{k=1}^H \sum_{x_h \in X_{i,h}} \sum_{a_h \in A_i} 1 \left\{ x_h \text{ visited, } a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h \right\} r_{i,k} \right]$$

$$= (\mathbb{E}_{\phi \pi_i} - \mathbb{E}_{\pi}) \left[ \sum_{x_h \in X_{i,h}} \sum_{a_h \in A_i} 1 \left\{ x_h \text{ visited, } a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h \right\} r_{i,k} \right]$$

$$+ (\mathbb{E}_{\phi \pi_i} - \mathbb{E}_{\pi}) \left[ \sum_{x_H \in X_{1:H}} \sum_{a_H \in A_i} 1 \left\{ x_H \text{ visited, } a_{1:H} \text{ recomd., and } \phi(x_H, a_{1:H}) = a_H \right\} r_{i,k} \right]$$

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 $h \leq H$, then the probability of $x_H$ is visited and $a_{1:H}$ are recommended are the same under $(\phi \circ \pi_i) \times \pi_{-i}$ and $\pi_i$; (ii) the randomness of $\sum_{h=1}^H r_{i,h}$ 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 $(x_h, a_h) \in X_{i,h} \times A_i$, $\phi \in \Phi_i^1$ and $\pi_i \in \Pi_i$, we define the trigger policy $\text{trig}(\pi_i, \phi \circ \pi_i), (x_h, a_h))$ to be a policy that plays $\pi_i$ before triggered by $(x_h, a_h)$ and plays $\phi \circ \pi_i$ 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 $(\phi \circ \pi_i) \times \pi_{-i}$ and $\text{trig}(\pi_i, \phi \circ \pi_i), (x_h, a_h)) \times \pi_{-i}$, which gives

$$\mathbb{E}_{\phi \pi_i} \left[ 1 \left\{ x_h \text{ visited, } a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h \right\} \right]$$

$$= \mathbb{E}_{\text{trig}(\pi_i, \phi \circ \pi_i), (x_h, a_h)) \times \pi_{-i}} \left[ 1 \left\{ x_h \text{ visited, } a_{1:h} \text{ recomd., and } \phi(x_h, a_{1:h}) \neq a_h \right\} \right].$$  

(18)
Consequently, we have
\[ \mathbb{E}_{\pi \sim \pi}(V_i^{(\phi \pi_i) \times \pi_{-i}} - V_i^\pi) \]
\[ = \sum_{h=1}^H \sum_{x_h \in X_{i,h}} \sum_{a_h \in A_i} \mathbb{E}_{\pi \sim \pi}(E_{(\phi \pi_i) \times \pi_{-i}} - \mathbb{E}_x) \left[ \mathbf{1}\{x_h \text{ visited, } a_{1,h} \text{ recomd., } \phi(x_h, a_{1,h}) \neq a_h\} \right] \]
\[ \leq \sum_{h=1}^H \sum_{x_h \in X_{i,h}} \sum_{a_h \in A_i} \mathbf{1}\{\phi(x_h, a_{1,h}) \neq a_h\} \times \mathbb{E}_{\pi \sim \pi}(E_{\text{trig}(\pi_i, (\phi \pi_i), (x_h, a_h)) \times \pi_{-i}} - \mathbb{E}_x) \left[ \mathbf{1}\{x_h \text{ visited and } a_{1,h} \text{ recomd.}\} \right] \]
\[ \leq \sum_{h=1}^H \sum_{x_h \in X_{i,h}} \sum_{a_h \in A_i} \text{TriggerGap}(\pi_i) = X_i A_i \cdot \text{TriggerGap}(\pi_i). \]

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 \( \tilde{\pi}_i(x_{i,H}) = a \) so that \( \text{trig}(\pi_i, \tilde{\pi}_i, (x_i, a)) = \pi_i \)); in (ii) we use the fact that \( \text{trig}(\pi_i, (\phi \pi_i), (x_i, a_h)) \) and \( \pi \) are identical on any infoset \( x \) such that \( x \neq x_h \) and \( x \neq (x_h, a_h) \), so that
\[ \left( \mathbb{E}_{\text{trig}(\pi_i, (\phi \pi_i), (x_h, a_h)) \times \pi_{-i}} - \mathbb{E}_x \right) \left[ \mathbf{1}\{a_{1,h} \text{ are not recommended or } x_h \text{ is not visited}\} \right] = 0. \]

Finally, take supremum over \( \phi \in \Phi_i^K \) and then take supremum over \( i \in [m] \), we get
\[ 1-\text{EFCEGap}(\pi_i) = \max_{i \in [m]} \max_{\phi \in \Phi_i^K} \mathbb{E}_{\pi \sim \pi}(V_i^{(\phi \pi_i) \times \pi_{-i}} - V_i^\pi) \leq \max_{i \in [m]} X_i A_i \cdot \text{TriggerGap}(\pi_i). \]

This proves the lemma. \( \square \)

C.2 Proof of Proposition C.2

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

Proof of \( K \)-EFCEGap(\( \pi_i \)) \( \leq (K + 1) \)-EFCEGap(\( \pi_i \)) 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 \( \phi \circ \pi_i \) and \( \phi^* \circ \pi_i \) gives the same policy. Given this claim, we have
\[ \max_{\phi \in \Phi_i^K} \mathbb{E}_{\pi \sim \pi}(V_i^{(\phi \pi_i) \times \pi_{-i}}) \leq \max_{\phi \in \Phi_i^{K+1}} \mathbb{E}_{\pi \sim \pi}(V_i^{(\phi \pi_i) \times \pi_{-i}}). \]

This implies \( K \)-EFCEGap(\( \pi_i \)) \( \leq (K + 1) \)-EFCEGap(\( \pi_i \)) 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 \( (x_{i,h}, b_{1,h-1}) \in \Omega_i^{(1),K} \cap \Omega_i^{(1),K+1} \) and any action \( a_h \in A_i \), 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 \( (x_{i,h}, b_{1,h-1}) \in \Omega_i^{(1),K+1} \cap \Omega_i^{(1),K} \) and any action \( a_h \in A_i \), we set \( \phi^*(x_{i,h}, b_{1,h-1}, a_h) = \phi^*(x_{i,h}, b_{1,h-1}, a_h) \), where \( h^{**} = \inf\{k \leq h : \sum_{k'=1}^k 1\{a_{h'} \neq b_{h'}\} = K\} \); (3) For any rechistory \( (x_{i,h}, b_{1,h'}) \in \Omega_i^{(1),K+1} \), we set \( \phi^*(x_{i,h}, b_{1,h'}) = \phi^*(x_{i,h}, b_{1,h'}) \) where \( h^{**} = \inf\{l \leq h' : \sum_{k'=1}^k 1\{a_{h'} \neq b_{h'}\} = K\} \). It is easy to see that for any \( \pi_i \in \Pi_i \), we have that \( \phi^* \circ \pi_i \) is the same as \( \phi \circ \pi_i \). This proves the claim.

Example of a game and a \( \pi \) with \( K \)-EFCEGap(\( \pi_i \)) \( = 0 \) but \( (K + 1) \)-EFCEGap(\( \pi_i \)) \( \geq 1/3 \)

For any \( K \geq 0 \), we consider a two-player game with \( H = K + 1 \) steps and perfect information. The action spaces are \( A_1 = \{1, 2\} \) for the first player and \( A_2 = \{1, 2\} \) for the second player in each time.
We next prove the second claim (converse bound does not hold), by constructing the following proposition.

We then 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 $\Phi_{i}^{K+1}$), she can receive reward 1 so that her received value is 1. As a consequence, we have $(K + 1)$-EFCEGap$(\pi) \geq 1 - 1/2 > 1/3$.

We then consider the $K$-EFCEGap. If the first player chooses to deviate at any step, she needs to play a different action from the second player at all time steps to receive an reward. 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 needs 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$(\pi) \leq 1/2 - 1/2 = 0$. This finishes the proof of the proposition.

C.3 Proof of Proposition C.3

(a) We first show that $\infty$-EFCEGap$(\pi) \leq \text{NFCEGap}(\pi)$. Indeed, for any $\infty$-EFCE strategy modification $\phi \in \Phi_{i}^{\infty}$, we let $\phi' \in \Phi_{i}^{\text{NFCE}}$ such that for any $x_{i,h} \in X_{i}$ and $\pi_{i} \in \Pi_{i}^{\text{pure}}$, $\phi'(x_{i,h}, \pi_{i}) := \phi(x_{i,h}, b_{i,h-1})$ where $b_{i,h-1} := \text{argmax}_{a \in A_{i}} \pi_{i}(a|x_{i,h-1})$ and $x_{i,1} \preceq \cdots \preceq x_{i,h-1} \preceq x_{i,h}$ are the unique history of infosets leading to $x_{i,h}$. By comparing the execution of $\phi \circ \pi$ (cf. Algorithm 1) and $\phi' \circ \pi_{i}$ (cf. Definition C.1), the policy $\phi \circ \pi_{i}$ exactly implements (i.e. is the same as) $\phi' \circ \pi_{i}$. This gives

$$\infty$$-EFCEGap$(\pi) = \max_{\phi \in \Phi_{i}^{\infty}} \mathbb{E}_{\pi_{i} \sim \pi_{i}} V_{i}^{(\phi \circ \pi_{i}) \times \pi_{i-1}} \leq \max_{\phi \in \Phi_{i}^{\text{NFCE}}} \mathbb{E}_{\pi_{i} \sim \pi_{i}} V_{i}^{(\phi \circ \pi_{i}) \times \pi_{i-1}} = \text{NFCEGap}(\pi).$

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

Example 1 (There exists an $\infty$-EFCE which is not $\varepsilon$-approximate NFCE with $\varepsilon = 1/20$): We consider a two-player game with 2 steps and perfect information. The set of infosets $X_{i,h}$ for both players gives $X_{1,1} = X_{2,1} = \{s_{0}\}$ and $X_{1,2} = X_{2,2} = \{s_{1,1}, s_{1,2}, s_{2,1}, s_{2,2}\}$. The action spaces are $A_{1} = \{a_{1}, a_{2}\}$ and $A_{2} = \{b_{1}, b_{2}\}$. If action pair $(a, b)$ $(i, j = 1, 2)$ is chosen at $s_{0}$, $s_{i,j}$ would be reached with probability 1. The reward for the first 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.
We first consider NFCE strategy modifications. On one hand, the first player only has motivation to modify his action at \( s_0 \) 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 \( \pi_i \) 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

\[
\text{NFCEGap}(\pi) \geq 1/16 > 1/20.
\]

On the other hand, for any \( \infty \)-EFCE strategy modification, taking \( a_2 \) at \( s_0 \) always has utility \( 1/2 \) since the actions taken by the second player at \( X_2 \) 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 \( \infty \)-EFCE strategy modification of \( \pi \) has value function \( 1/2 \), so does \( \pi \). Consequently, \( \pi \) is an exact \( \infty \)-EFCE, i.e. \( \infty \)-EFCEGap(\( \pi \)) = 0.

(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 \( \pi_i \in \Pi_i^{\text{pure}} \), \( \phi \circ \pi_i \) takes action \( \phi(x_i, \emptyset) \) at \( x_i \). So \( \phi \) is equivalent to a modification \( \phi' \in \Phi_i^{\text{NFCCE}} \) which satisfies \( \phi'(x_i) = \phi(x_i, \emptyset) \) for all \( x_i \in X_i \). Here, the equivalence means that \( \phi \circ \pi_i = \phi' \circ \pi_i \) for any \( \pi_i \in \Pi_i^{\text{pure}} \). This gives

\[
0\text{-EFCEGap}(\pi) = \max_{\phi \in \Phi_i^0} \mathbb{E}_{\pi \sim \pi} V_i(\phi \circ \pi_i) = \max_{\phi \in \Phi_i^{\text{NFCCE}}} \mathbb{E}_{\pi \sim \pi} V_i(\phi \circ \pi_i) = \text{NFCCEGap}(\pi),
\]

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 \( \mu^\phi \) as follows: For any \( a_k \in A_i \),

\[
\mu^\phi_k(a_k|x_{i,h}, b_{1:h-1}, b_h) := 1 \{a_k = \phi(x_{i,h}, b_{1:h-1}, b_h)\} \quad \text{for all } (x_{i,h}, b_{1:h-1}, b_h) \in \Omega_i^{(1), K} \times \mathcal{A}_i,
\]

\[
\mu^\phi_k(a_k|x_{i,h}, b_{1:h'}) := 1 \{a_k = \phi(x_{i,h}, b_{1:h'})\} \quad \text{for all } (x_{i,h}, b_{1:h'}) \in \Omega_i^{(II), K}.
\]

In words, \( \mu^\phi_k(\cdot|x_{i,h}, b_{1:h-1}, b_h) \in \Delta(\mathcal{A}_i) \) 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 \( b_h \in \mathcal{A}_i \); \( \mu^\phi_k(\cdot|x_{i,h}, b_{1:h'}) \) 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

\[
\phi_k(\cdot|x_{i,h}, b_{1:h}) := \mu^\phi_k(\cdot|x_{i,h}, b_{1:h-1}, b_h), \quad \phi_k(\cdot|x_{i,h}, b_{1:h'}) := \mu^\phi_k(\cdot|x_{i,h}, b_{1:h'}).
\]

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

\[
D(a_{1:k}, b_{1:k}) := \sum_{k=1}^{K} \{a_k \neq b_k\},
\]

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

\[
\delta^{\leq K-1}(a_{1:h-1}, b_{1:h-1}) := 1 \{D(a_{1:h-1}, b_{1:h-1}) \leq K - 1\};
\]

\[
\delta^K(a_{1:h-1}, b_{1:h-1}) := 1 \{D(a_{1:h-1}, b_{1:h-1}) = K\}.
\]

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

\[
(\phi \circ \pi_i)_1^{h}(x_h, a_h) = \sum_{b_{1:h}} \delta^{\leq K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi_k(a_k|x_k, b_{1:k}) \prod_{k=1}^{h} \pi_i(b_k|x_k).
\]
+ \sum_{b_1:h} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi_k(a_k \vert x_k, b_{1:k \wedge \tau_K}) \prod_{k=1}^{\tau_K} \pi_i(b_k \vert x_k),

where

\tau_K := \inf \left\{ h' \leq h - 1 : \sum_{h''=1}^{h'} 1 \{ a_{h''} \neq b_{h''} \} \geq K \right\}

(19)

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

\[
\sum_{(x_h, a_h) \in X_{i,h} \times A_i} \frac{(\phi \circ \pi_i)_{1:h}(x_h, a_h)}{\pi_i^{*_{1:h}}(x_h, a_h)} = X_{i,h} A_i.
\]

**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 \(a_1, \ldots, a_{h-1}\). The sequence-form expression \((\phi \circ \pi_i)(x_h, a_h)\) is the probability of \((\phi \circ \pi_i)\) choosing \(a_k\) 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 \(\{|h' \in [h-1] : a_{h'} \neq b_{h'}\}| \leq K - 1\), the conditional probability of \((\phi \circ \pi)\) choosing \(a_k\) at \(x_k\) for all \(k \in [h]\) is \(\prod_{k=1}^{h} \phi_k(a_k \vert x_k, b_{1:k})\), as the player would always swap the action; If \(\{|h' \in [h] : a_{h'} \neq b_{h'}\}| \geq K\), the conditional probability of \((\phi \circ \pi)\) choosing \(a_k\) at \(x_k\) for all \(k \in [h]\) is \(\prod_{k=1}^{h} \phi_k(a_k \vert x_k, b_{1:k \wedge \tau_K})\). So by the law of total probability, we have

\[
(\phi \circ \pi_i)_{1:h}(x_h, a_h) = \sum_{b_1:h} \delta^{\leq K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi_k(a_k \vert x_k, b_{1:k}) \prod_{k=1}^{h} \pi_i(b_k \vert x_k)
\]

\[
+ \sum_{b_1:h} 1 \{ \{|h' \in [h] : a_{h'} \neq b_{h'}\}| \geq K \} \prod_{k=1}^{h} \phi_k(a_k \vert x_k, b_{1:k \wedge \tau_K}) \prod_{k=1}^{h} \pi_i(b_k \vert x_k).
\]

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

\[
\sum_{b_1:h} 1 \{ \{|h' \in [h] : a_{h'} \neq b_{h'}\}| \geq K \} \prod_{k=1}^{h} \phi_k(a_k \vert x_k, b_{1:k \wedge \tau_K}) \prod_{k=1}^{h} \pi_i(b_k \vert x_k)
\]

\[
= \sum_{(h', b_{1:h'}) : \sum_{k=1}^{h'} 1(a_k \neq b_k) = \tau_K \wedge b_{k':0} \neq a_{k':0}} \prod_{k=1}^{h'} \phi_k(a_k \vert x_k, b_{1:k \wedge \tau_K}) \prod_{k=1}^{h'} \pi_i(b_k \vert x_k).
\]

The last equality is because we can append \(a_{h'+2}\) 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. \[\square\]

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 \(x_h \in X_{i,h}\) with \((a_1, \ldots, a_{h-1})\) 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 \((\phi \circ \pi_i) \times \pi_{-i}\) is

\[
P_{\phi \circ \pi_i \times \pi_{-i}}(x_h \text{ is reached by the } i\text{th player}) = \sum_{b_{1:h-1}} \delta^{\leq K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k \vert x_k, b_{1:k}) \cdot \prod_{k=1}^{h-1} \pi_i(b_k \vert x_k) p_{1:h}^{\pi_{-i}}(x_{i,h})
\]
\[
\begin{aligned}
&+ \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k} \land \tau_K) \cdot \prod_{k=1}^{r_K} \pi_i(b_k|x_k)p_{1:h}^{\pi_i}(x_{1:h}).
\end{aligned}
\]

**Proof.** We have
\[
\mathbb{P}_{x, \pi_i \times \pi_{-i}}(x_h \text{ is reached by the } i^{th} \text{ player}) = \sum_{a \in A_i} (\phi \circ \pi_i)(x_h, a)p_{1:h}^{\pi_i}(x_{1:h}),
\]
where \(p_{1:h}^{\pi_i}(x_{1:h})\) is defined in equation (12). So applying Lemma D.1 yields the desired result. \(\square\)

As each step \(h\), one (and only one) \(x_h \in X_{i,h}\) 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
\[
\sum_{x_h \in X_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \cdot \prod_{k=1}^{h-1} \pi_i(b_k|x_k)p_{1:h}^{\pi_i}(x_{1:h})
\]
\[
+ \sum_{x_h \in X_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k} \land \tau_K) \cdot \prod_{k=1}^{r_K} \pi_i(b_k|x_k)p_{1:h}^{\pi_i}(x_{1:h}) = 1.
\]

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

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

**Proof.** This follows directly by the definition of \(K\)-EFCEGap:
\[
K\text{-EFCEGap}(\pi) = \max_{i \in [m]} \max_{\phi \in \Phi_i^K} \left( V_i^{\phi \pi_i \times \pi_{-i}} - V_i^{\pi} \right)
\]
\[
= \max_{i \in [m]} \max_{\phi \in \Phi_i^K} \mathbb{E}_{\pi \sim \pi} \left[ V_i^{\phi \pi_i \times \pi_{-i}} - V_i^{\pi} \right]
\]
\[
= \max_{i \in [m]} \max_{\phi \in \Phi_i^K} \frac{1}{T} \sum_{t=1}^T \left( V_i^{\phi \pi_t^i \times \pi_{-i}} - V_i^{\pi_t^i} \right)
\]
\[
= \max_{i \in [m]} R_{i,K}^T / T.
\]

This proves the lemma. \(\square\)

For \((x_{i,h}, b_{1:h-1}, \varphi) \in \Omega_{i,K} \times \Phi_i^K\), we define the immediate local swap regret as
\[
\hat{R}_{(x_{i,h}, b_{1:h-1}, \varphi)}^T := \sum_{t=1}^T \prod_{k=1}^{h-1} \pi_i^t(b_k|x_k) \left( \pi_i^t(x_{1:h} | \cdot) - \varphi \circ \pi_i^t(x_{1:h}, \cdot) \right),
\]

28
and the (overall) local swap regret as
\[
\hat{R}_{(x_h,b_{1:h-1})}^{T,\text{swap}} := \max_{\phi} \hat{R}_{(x_h,b_{1:h-1}),\phi}^{T}.
\]

For \((x_i,h,b_{1:h'},a) \in \Omega_{x_i}^{(1)},K \times A_i\), we define the immediate local external regret as
\[
\hat{R}^{T}_{(x_h,b_{1:h',a}),a} := \sum_{t=1}^{T} \prod_{i,k=1}^{H'} \pi_i^k(b_k|x_k) \left( \langle \pi_i^k(h \cdot |x_{i,h}), L_i^t(x_{i,h},\cdot) \rangle - L_i^t(x_{i,h},a) \right),
\]
and the (overall) local external regret as
\[
\hat{R}^{T,\text{ext}}_{(x_h,b_{1:h',a}),a} := \max_{a \in A_i} \hat{R}^{T}_{(x_h,b_{1:h',a}),a}.
\]

\textbf{K-EFCE regret decomposition} Our main result in this section is the following regret decomposition that decomposes the K-EFCE regret \(R_{i,K}^{T}\) into combinations of local regrets at each rehistory.

\textbf{Lemma E.2} (Regret decomposition for K-EFCE regret). We have \(R_{i,K}^{T} \leq \sum_{h=1}^{H} R_{i}^{T}\), with
\[
R_{i}^{T} := \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_{i}^{T,\text{swap}}(x_h;\phi) + \max_{\phi \in \Phi_i^K} \sum_{x_h \in \mathcal{X}_{i,h}} G_{i}^{T,\text{ext}}(x_h;\phi),
\]
where
\[
G_{i}^{T,\text{swap}}(x_h;\phi) := \sum_{b_{1:h-1}} \delta^{<K}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k}) \hat{R}_{(x_h,b_{1:h-1})}^{T,\text{swap}}.
\]
and
\[
G_{i}^{T,\text{ext}}(x_h;\phi) := \sum_{b_{1:h-1}} \delta^{K}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k\wedge \tau_K}) \hat{R}_{(x_h,b_{1:1})}^{T,\text{ext}} = \sum_{h'=K}^{h} \sum_{b_{1:h'}} \mathbf{1}_{\tau_K = h'} \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k\wedge h'}) \hat{R}_{(x_h,b_{1:h'})}^{T,\text{ext}}.
\]

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

\textbf{Proof of Lemma E.2}. We begin by performing the following performance decomposition
\[
R_{i,K}^{T} = \max_{\phi \in \Phi_i^K} \sum_{t=1}^{T} (V_{i \circ \pi_i^{t-1}} - V_{i}^{t})
\]
\[
= \max_{\phi \in \Phi_i^K} \sum_{t=1}^{T} \left( E_{\phi \circ \pi_i^{t-1}} \left[ \sum_{h=1}^{H} \hat{R}_{i,h}^{t} \right] - E_{\phi} \left[ \sum_{h=1}^{H} \hat{R}_{i,h}^{t} \right] \right)
\]
\[
= \max_{\phi \in \Phi_i^K} \sum_{h=1}^{H} \sum_{t=1}^{T} \left( (\phi \circ \pi_i^{t-1})_{1:h,\pi_i^{t-1}_{h+1:H}} \times \pi_i^{t-1} \right) \left[ \sum_{t=1}^{T} \hat{R}_{i,h}^{t} \right] - E_{\phi} \left[ (\phi \circ \pi_i^{t-1})_{1:h,\pi_i^{t-1}_{h+1:H}} \times \pi_i^{t-1} \right] \left[ \sum_{t=1}^{T} \hat{R}_{i,h}^{t} \right]
\]
\[
= \max_{\phi \in \Phi_i^K} \sum_{h=1}^{H} \sum_{t=1}^{T} \left( (\phi \circ \pi_i^{t-1})_{1:h,\pi_i^{t-1}_{h+1:H}} \times \pi_i^{t-1} \right) \left[ \sum_{t=1}^{T} (1 - r_{i,k}) \right] - E_{\phi} \left[ (\phi \circ \pi_i^{t-1})_{1:h,\pi_i^{t-1}_{h+1:H}} \times \pi_i^{t-1} \right] \left[ \sum_{t=1}^{T} (1 - r_{i,k}) \right].
\]
Here, \(((\phi \circ \pi_i^{t-1})_{1:h,\pi_i^{t-1}_{h+1:H}} \times \pi_i^{t-1})\) refers to the policy that the \(i\)th player uses \(\phi \circ \pi_i^{t-1}\) for the first \(h\) step, and then uses \(\pi_i^{t-1}\) where as other players always use \(\pi_i^{t-1}\). The last step use the fact that \(((\phi \circ \pi_i^{t-1})_{1:h,\pi_i^{t-1}_{h+1:H}} \times \pi_i^{t-1})\) and \(((\phi \circ \pi_i^{t-1})_{1:h-1,\pi_i^{t-1}_{h:H}} \times \pi_i^{t-1})\) are the same for the first \(h-1\) steps, so the expected reward in first \(h-1\) steps are the same, too. Therefore, define
\[
\hat{R}_{i}^{T} = \max_{\phi \in \Phi_i^K} \sum_{t=1}^{T} \left( (\phi \circ \pi_i^{t-1})_{1:h-1,\pi_i^{t-1}_{h:H}} \times \pi_i^{t-1} \right) \left[ \sum_{t=1}^{T} (1 - r_{i,k}) \right] - E_{\phi} \left[ (\phi \circ \pi_i^{t-1})_{1:h-1,\pi_i^{t-1}_{h:H}} \times \pi_i^{t-1} \right] \left[ \sum_{t=1}^{T} (1 - r_{i,k}) \right],
\]
we have $R_{1,K}^T \leq \sum_{h=1}^H \tilde{R}_h^T$.

We next show that

$$
\tilde{R}_h^T \leq \max_{\phi \in \Phi_h^T} \sum_{x_h \in X_{1,h}} G_{h}^{T,\text{swap}}(x_h; \phi) + \max_{\phi \in \Phi_h^T} \sum_{x_h \in X_{1,h}} G_{h}^{T,\text{ext}}(x_h; \phi) = R_h^T,
$$

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

$$
\mathbb{E}_{((\phi \circ \pi_i)_{1:h-1} \pi_i^T, h: H)} \left[ \sum_{k=h}^H (1 - r_{i,k}) \right] = \left( \prod_{k=1}^{h-1} \pi_i^T(b_k|x_k) \prod_{k=1}^{h-1} \phi_h(a_k|x_k, b_{1:k-1} \wedge \tau_K) \right) \left( \sum_{k=1}^H (1 - r_{i,k}) \right)
$$

where we assume $\tau_K := \inf\{h' \leq h-1 : \sum_{h''=1}^{h'} (a_{h''} \neq b_{h''}) \geq K\}$. Similarly,

$$
\mathbb{E}_{((\phi \circ \pi_i)_{1:h-1} \pi_i^T, h: H)} \left[ \sum_{k=h}^H (1 - r_{i,k}) \right] = \left( \prod_{k=1}^{h-1} \pi_i^T(b_k|x_k) \prod_{k=1}^{h-1} \phi_h(a_k|x_k, b_{1:k-1} \wedge \tau_K) \right) \left( \sum_{k=1}^H (1 - r_{i,k}) \right)
$$

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

$$
\tilde{R}_h^T = \max_{\phi \in \Phi_h^T} \sum_{t=1}^T \left( \mathbb{E}_{((\phi \circ \pi_i)_{1:h-1} \pi_i^T, h: H)} \left[ \sum_{k=h}^H (1 - r_{i,k}) \right] - \mathbb{E}_{((\phi \circ \pi_i)_{1:h-1} \pi_i^T, h: H)} \left[ \sum_{k=h}^H (1 - r_{i,k}) \right] \right)
$$

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

$$
\tilde{R}_h^T = \max_{\phi \in \Phi_h^T} \sum_{x_h \in X_{1,h}, b_{1:h-1}} \left( \prod_{k=1}^{h-1} \phi_h(a_k|x_k, b_{1:k-1} \wedge \tau_K) \sum_{t=1}^T \sum_{k=1}^{h-1} \pi_i^T(b_k|x_k) \left( \pi_i^T(\cdot|x_h) - \phi_h(\cdot|x_h, b_{1:h-1} \wedge \tau_K), L_{i,h}(x_h, \cdot) \right) \right)
$$

where

$$
I_h := \sum_{x_h \in X_{1,h}, b_{1:h-1}} \delta^{\leq K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_h(a_k|x_k, b_{1:k})
$$

and

$$
\Pi_h := \sum_{x_h \in X_{1,h}, b_{1:h-1}} \delta^{\leq K-1}(D(a_{1:h-1}, b_{1:h-1}) \geq K) \prod_{k=1}^{h-1} \phi_h(a_k|x_k, b_{1:k-1} \wedge \tau_K)
$$
Above, (i) follows by the definition of the local swap regret in (21). For times at some step \( D_{\text{max}} \) over all \( \phi \) \( K \)

\[
\max_{\phi \in \Phi_t^K} I_h = \max_{\phi \in \Phi_t^K} \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \sum \...
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 $\sum_{h'=h+1:h'-1} \prod_{k=h+1}^{h'-1} \pi_k'(b_k|x_k) = 1$; (ii) follows by definition of the local external regret in (22). Furthermore, for fixed $x_h$, we can expand $b_{1:h'}$ to $b_{1:h} := (b_{1:h'}, a_{h'+1}, \ldots, a_{h-1})$ such that $D(a_{1:h'-1}, b_{1:h'-1}) = K$. Then we can rewrite

$$\sum_{h'=h+1}^{h-1} \sum_{b_{1:h'}} 1 \{ \tau_K = h' \} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k:h'}) \hat{R}^{T,ext}_{(x_h, b_{1:h'}, x_h)}$$

as

$$\sum_{h'=h}^{h-1} \sum_{b_{1:h'}} 1 \{ \tau_K = h', D(a_{1:h'-1}, b_{1:h'-1}) = K \} \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k:h'}) \hat{R}^{T,ext}_{(x_h, b_{1:h'}, x_h)}$$

Finally, combining the above bounds for $\max_{\phi \in \Phi^K} \Pi_h$ and $\max_{\phi \in \Phi^K} \Pi_h$ gives the desired result:

$$\bar{R}_h^T = \max_{\phi \in \Phi^K} (I_h + \Pi_h) \leq \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} G_h^{T,swap}(x_h) + \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} G_h^{T,ext}(x_h) = R_h^T.$$  

\qed

## F Proofs for Section 4

This section is devoted to proving Theorem 3.

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

$$R_{i,K}^T = \max_{\phi \in \Phi^K} \sum_{t=1}^{T} (V_{i}^\phi \pi^\phi_t \times \pi^\phi_t - V_{i}^\phi')$$

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

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

$$R_h^T := \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} G_h^{T,swap}(x_h; \phi) + \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} G_h^{T,ext}(x_h; \phi).$$

The following two lemmas bound two terms

$$\sum_{h=1}^{H} \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} G_h^{T,swap}(x_h; \phi) \text{ and } \sum_{h=1}^{H} \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} G_h^{T,ext}(x_h; \phi).$$

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

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

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

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

$$\sum_{h=1}^{H} \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} G_h^{T,swap}(x_h; \phi) \leq \sqrt{H^4 \left( \frac{H}{K \wedge H} \right) X_i A_i^{K \wedge H} T \log A_i}.$$
We remark that the above proof does not depend on the particular choice of \( x \). Then we have
\[
\sum_{h=1}^{H} \max_{\phi \in \Phi^h} \sum_{x_h \in X_{i,h}} G_{h,x}^{T,\text{ext}}(x_h; \phi) \leq O\left( \sqrt{H^4 \left( \frac{H}{K \land H} \right) X_i A_{i,K \land H,T} \log A_i} \right). \tag{25}
\]
Combining Lemma F.1 & F.2, we obtain regret bound
\[
R_{i,K}^T \leq \sum_{h=1}^{H} R_h^T \leq \sum_{h=1}^{H} \left( \max_{\phi \in \Phi^h} \sum_{x_h \in X_{i,h}} G_{h,x}^{T,\text{swap}}(x_h; \phi) + \max_{\phi \in \Phi^h} \sum_{x_h \in X_{i,h}} G_{h,x}^{T,\text{ext}}(x_h; \phi) \right)
= O\left( H^4 \left( \frac{H}{K \land H} \right) A_{i,K \land H,T} X_i T \log A_i \right). \tag{26}
\]
In particular, as long as
\[
T \geq O\left( H^4 \left( \frac{H}{K \land H} \right) \left( \max_{i \in [m]} X_i A_{i,K \land H} \right) \log A_i / \epsilon^2 \right),
\]
we have by the online-to-batch lemma (Lemma E.1) that the average policy \( \pi = \text{Unif}\{\pi_1^T, \ldots, \pi_m^T\} \) satisfies
\[
K\text{-EFCEGap}(\pi) = \frac{\max_{i \in [m]} R_{i,K}^T}{T} \leq \max_{i \in [m]} O\left( H^4 \left( \frac{H}{K \land H} \right) \left( \max_{i \in [m]} X_i A_{i,K \land H} \right) \log A_i / T \leq \epsilon. \right.) \tag{27}
\]
This proves Theorem 3.

We remark that the above proof does not depend on the particular choice of \( \pi_1^T \), and thus the regret bound (25) also holds even if we control the \( i \)-th player only, and \( \pi_{-i}^T \) are arbitrary (potentially adversarial depending on all information before iteration \( t \) starts). This directly gives the following corollary.

**Corollary F.1** \((K\text{-EFCE}) \) regret bound for \( K\text{-EFR} \) against adversarial opponents. For any \( 0 \leq K \leq \infty, \epsilon \in (0, H) \), suppose the \( i \)-th player runs Algorithm 2 together against arbitrary (potentially adversarial) opponents \( \{\pi_{i,h}^T\}_{h=1}^T \), where \( \text{REGALG} \) is instantiated as Algorithm 4 with learning rates specified in (8). Then the \( i \)-th player achieves \( K\text{-EFCE} \) regret bound:
\[
R_{i,K}^T = \max_{\phi \in \Phi^h} \sum_{t=1}^{T} V_i^t - V_i^{\pi_1^T, \ldots, \pi_{i-1}^T, \pi_i^T} \leq O\left( H^4 \left( \frac{H}{K \land H} \right) X_i A_{i,K \land H,T} \log A_i \right).
\]

\section{Proof of Lemma F.1}

**Proof.** Recall that \( G_{h,x}^{T,\text{swap}}(x_h; \phi) \) is defined as
\[
G_{h,x}^{T,\text{swap}}(x_h; \phi) := \sum_{b_{1:h-1}} \delta_{\leq K}^{-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \tilde{R}_{(x_h,b_{1:h-1})}^{T,\text{swap}}(x_h,b_{1:h-1}),
\]
where for each \( h \in [H] \) and \( (x_h, b_{1:h-1}) \in \Omega_t^{(t),K} \),
\[
\tilde{R}_{(x_h,b_{1:h-1})}^{T,\text{swap}} = \max_{\phi} \sum_{t=1}^{T} \prod_{h=1}^{h-1} \pi_{i,h}^t(b_k|x_k) \left( \pi_{i,h}^t(x_h,h) - \phi \pi_{i,h}^t(x_h,h) \right),
\]
For \( x_h \in X_{i,h} \), we first apply regret minimization lemma (Lemma A.2) on \( R_{x_h} \) to give an upper bound on \( \tilde{R}_{(x_h,b_{1:h-1})}^{T,\text{swap}} \). Recall in Algorithm 2 with \( \text{REGALG} \) instantiated as Algorithm 4, each \( R_{x_h} \)
observes time selection functions $S^{t}_{b_{1:h-1}} = \prod_{k=1}^{h-1} \pi^{t}_{i}(b_{k}|x_{k})$ for $(x_{h}, b_{1:h-1}) \in \Omega^{(1)}_{i,K}$ and loss vector $L^{t}_{i,h}(x_{h}, \cdot)$. Suppose $\mathcal{R}_{x_{h}}$ uses learning rate $\eta_{x_{h}} = \eta$ for all $x_{h} \in \mathcal{X}_{h}$. Then, by the regret bound with respect to a time selection function index and strategy modification pair (Lemma A.2), we have
\[
\max_{\varphi} \sum_{t=1}^{T} \sum_{h=1}^{h-1} \pi^{t}_{i}(b_{h}|x_{h}) \left( \langle \pi^{t}_{i,h}(\cdot|x_{i,h}), \varphi \circ \pi^{t}_{i,h}(\cdot|x_{i,h}) \rangle, L^{t}_{i,h}(x_{i,h}, \cdot) \right) \\
\leq \eta \sum_{t=1}^{T} \sum_{h=1}^{h-1} \pi^{t}_{i}(b_{h}|x_{h}) \|L^{t}_{i,h}(x_{h}, \cdot)\|_{\infty} \langle \pi^{t}_{i}(\cdot|x_{h}), L^{t}_{i,h}(x_{h}, \cdot) \rangle + \frac{(A_{i} + H) \log A_{i}}{\eta} \\
\leq \eta H^{2} \sum_{t=1}^{T} \sum_{h=1}^{h-1} \pi^{t}_{i}(b_{h}|x_{h}) \pi^{t}_{i}(x_{h}) + \frac{(A_{i} + H) \log A_{i}}{\eta}.
\]
Above, we used (i) $\|L^{t}_{i,h}(x_{h}, \cdot)\|_{\infty} \leq H p^{t}_{i,h}(x_{h})$, and (ii) $(\|B^{*} + |B^{*}|) |\Psi^{*}| \leq A_{i}^{H + A_{i}}$ since the number of rechostories is no more than $A_{i}^{H}$. Then we can get
\[
\max_{\Phi \in \Phi^{K}_{i}} \sum_{x_{h} \in \mathcal{X}_{i,h}} G_{h}^{T,\text{swap}}(x_{h}; \Phi) \\
= \max_{\Phi \in \Phi^{K}_{i}} \sum_{x_{h} \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k}) \\
\times \max_{\varphi} \sum_{t=1}^{T} \sum_{h=1}^{h-1} \pi^{t}_{i}(b_{h}|x_{h}) \left( L^{t}_{i,h}(x_{h}, \pi^{t}_{i}(x_{h})) - L^{t}_{i,h}(x_{h}, \varphi \circ \pi^{t}_{i}(x_{h})) \right) \\
\leq \eta H^{2} \max_{\Phi \in \Phi^{K}_{i}} \sum_{x_{h} \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k}) \sum_{t=1}^{T} \sum_{h=1}^{h-1} \pi^{t}_{i}(b_{h}|x_{h}) p^{t}_{i,h}(x_{h}) + \frac{(A_{i} + H) \log A_{i}}{\eta} \max_{\Phi \in \Phi^{K}_{i}} \sum_{x_{h} \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k}).
\]

Letting
\[
I_{h} := \max_{\Phi \in \Phi^{K}_{i}} \sum_{x_{h} \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k});
\]
\[
II_{h} := \max_{\Phi \in \Phi^{K}_{i}} \sum_{x_{h} \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k}) \sum_{t=1}^{T} \sum_{h=1}^{h-1} \pi^{t}_{i}(b_{h}|x_{h}) p^{t}_{i,h}(x_{h}).
\]

For fixed $\Phi \in \Phi^{K}_{i}$ and $x_{h} \in \mathcal{X}_{i,h}$, by counting the number of $b_{1:h-1}$ such that $D(a_{1:h-1}, b_{1:h-1}) \leq K - 1$, we have
\[
\sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k}) \\
\leq \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \leq \left( \frac{h-1}{(K-1) \wedge (h-1)} \right) A_{i}^{(K-1) \wedge (h-1)}.
\]

Consequently,
\[
I_{h} \leq X_{i,h} \left( \frac{h-1}{(K-1) \wedge (h-1)} \right) A_{i}^{K \wedge h-1}. \tag{27}
\]

Note that by Corollary D.2, for fixed $t$,
\[
\sum_{x_{h} \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k}) \prod_{h=1}^{h-1} \pi^{t}_{h}(b_{h}|x_{h}) p^{t}_{i,h}(x_{i,h}) \leq 1.
\]

Consequently, we have
\[
\max_{\Phi \in \Phi^{K}_{i}} \sum_{x_{h} \in \mathcal{X}_{i,h}} \sum_{b_{1:h-1}} \delta^{\leq K-1} \pi_{i,h}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}|x_{k}, b_{1:k}) \sum_{t=1}^{T} \sum_{h=1}^{h-1} \pi^{t}_{i}(b_{h}|x_{h}) p^{t}_{i,h}(x_{i,h}) \leq T.
\]
This yields that

\[ \Pi_h \leq T. \]  

(28)

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

\[
\begin{align*}
\sum_{h=1}^{H} \max_{\phi \in \Phi_h} \sum_{x_h \in X_i, h} G_h^{T,\text{swap}}(x_h; \phi) & \leq \sum_{h=1}^{H} \frac{(A_i + H) \log A_i}{\eta} I_h + H^2 \eta \Pi_h \\
& \leq H^2 \eta T + \sum_{h=1}^{H} \frac{(A_i + H) \log A_i}{\eta} X_i h \left( H - 1 \right) \left( K - 1 \right) \left( h - 1 \right) \left( h - 1 \right) A_i K^{\wedge K - 1} \\
& \leq H^2 \eta T + \frac{(A_i + H) \log A_i}{\eta} X_i \left( H - 1 \right) \left( K \wedge H - 1 \right) A_i K^{\wedge K - 1} \\
& \leq H^2 \eta T + \frac{H \log A_i}{\eta} X_i \left( H \right) \left( K \wedge H \right) A_i K^{\wedge K - 1}. 
\end{align*}
\]

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

\[
\sum_{h=1}^{H} \max_{\phi \in \Phi_h} \sum_{x_h \in X_i, h} G_h^{T,\text{swap}}(x_h; \phi) \leq \sqrt{H^4 \left( H \right) \left( K \wedge H \right) X_i A_i K^{\wedge K - 1} T \log A_i}. 
\]

\( \square \)

### F.2 Proof of Lemma F.2

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

\[
G_h^{T,\text{ext}}(x_h; \phi) := \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k \wedge K}) \tilde{R}^{T,\text{ext}}(x_h, b_{1:K}).
\]

Since \( (x_h, b_{1:K}) \in \Omega_i^{(1), K} \), we have

\[
\tilde{R}^{T,\text{ext}}(x_h, b_{1:K}) = \max_a \sum_{a} \prod_{t=1}^{T} \pi_t(b_t | x_t) \left( \pi_t^i(x_t, h), L_t^i(x_t, h, \cdot) \right) \left( \pi_t^i(x_t, h), L_t^i(x_t, h, a) \right).
\]

For \( x_h \in X_i, h \), we can apply regret minimization lemma (Lemma A.2) on \( R_{x_h} \) give an upper bound on \( \tilde{R}^{T,\text{swap}}_{(x_h, b_{1:K})} \). Recall in Algorithm 2 with REGALG instantiated as Algorithm 4, each regret minimizer \( R_{x_h} \) observes time selection functions \( S_t^{i, b_{1:K}} = \prod_{k=1}^{b_{1:K}} \pi_t^i(b_t | x_t) \) for \( (x_h, b_{1:K}) \in \Omega_i^{(1), K} \) and losses \( L_t^i(x_h, \cdot) \). Suppose all \( R_{x_h} \) use the same learning rate \( \eta \), by the bound on regret with respect to a time selection function index and strategy modification pair (Lemma A.2), we have

\[
\max_a \sum_{t=1}^{T} \prod_{k=1}^{b_{1:K}} \pi_t^i(b_t | x_t) \left( \pi_t^i(x_t, h), L_t^i(x_t, h, \cdot) \right) - L_t^i(x_t, h, a) \leq \eta \sum_{t=1}^{T} \prod_{k=1}^{b_{1:K}} \pi_t^i(b_t | x_t) \||L_t^i(x_t, h, \cdot)\|_\infty \pi_t^i(x_t, h, \cdot) + \frac{\log((|B^e| + |B^e|)|\Psi^e|)}{\eta} \\
\leq \eta H^2 \sum_{t=1}^{T} \prod_{k=1}^{b_{1:K}} \pi_t^i(b_t | x_t) \pi_t^i(x_t, h) + \frac{2H \log A_i}{\eta}.
\]

Here, we use (i) \( |L_t^i(x_t, h, \cdot)\|_\infty \leq H p_{x_t, h}^e \) \( (x_h) \), and (ii) \( (|B^e| + |B^e|)|\Psi^e| \leq A_i H^1 \) since the number of recollections is no more than \( A_i H^1 \). Then we can get

\[
\max_{\phi \in \Phi_h} G_h^{T,\text{ext}}(x_h; \phi)
\]

35
\[= \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} \sum_{b_{h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K}) \times \max_{\eta} \sum_{t=1}^{\tau_K} \prod_{k=1}^{h} \pi_{x}^t (b_k | x_k) \left( (\pi_{x}^t (x_i, h), L_{i,h}^t(x_i, h, \cdot)) - L_{i,h}^t(x_i, h, a) \right) \]

\[\leq \eta H^2 \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} \sum_{b_{h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K}) \cdot \sum_{t=1}^{\tau_K} \prod_{k=1}^{h} \pi_{x}^t (b_k | x_k) \eta_i^{T} (x_h) + \frac{2H \log A_i}{\eta} \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} \sum_{b_{h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K}). \]

Similar to the proof of Lemma F.1, letting

\[I_h := \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} \sum_{b_{h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K});\]

\[II_h := \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} \sum_{b_{h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K}) \cdot \sum_{t=1}^{\tau_K} \prod_{k=1}^{h} \pi_{x}^t (b_k | x_k) \eta_i^{T} (x_h). \]

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

\[\sum_{b_{1:h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K}) \leq \sum_{b_{1:h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \leq (h - 1) A_i^{K\wedge (h-1)}. \]

Consequently,

\[I_h \leq X_{i,h} \left( \frac{h - 1}{K \wedge h} \right) A_i^{K\wedge (h-1)}. \tag{29} \]

Note that for fixed \(t\), by Corollary D.2,

\[\sum_{x_h \in X_{i,h}} \sum_{b_{h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K}) \prod_{k=1}^{h} \pi_{x}^t (b_k | x_k) \eta_i^{T} (x_i, h) \leq 1. \]

Consequently, we have

\[\max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}} \sum_{b_{h-1}} \delta^K (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k\land \tau_K}) \cdot \sum_{t=1}^{\tau_K} \prod_{k=1}^{h} \pi_{x}^t (b_k | x_k) \eta_i^{T} (x_i, h) \leq T. \]

This yields that

\[\Pi_h \leq T. \tag{30} \]

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

\[\sum_{h=1}^{H} \max_{\phi \in \Phi^K} G_h^T,ext (x_h; \phi) \leq \sum_{h=1}^{H} \left( \frac{2H \log A_i}{\eta} I_h + H^2 \eta I_h \right) \leq H^3 \eta T + \sum_{h=1}^{H} \frac{2H \log A_i}{\eta} X_{i,h} \left( \frac{h - 1}{K \wedge h} \right) A_i^{K\wedge H} \]

\[\leq H^3 \eta T + \frac{2H \log A_i}{\eta} X_{i,h} \left( \frac{H - 1}{K \wedge H} \right) A_i^{K\wedge H} \]

\[\leq H^3 \eta T + \frac{2H \log A_i}{\eta} X_{i,h} \left( \frac{H}{K \wedge H} \right) A_i^{K\wedge H}. \]

As we choose \(\eta = \sqrt{\frac{H}{(K \wedge H) A_i^{K\wedge H} X_{i,h} \log A_i}}\) per (8), we have

\[\sum_{h=1}^{H} \max_{\phi \in \Phi^K} G_h^T,ext (x_h; \phi) \leq 3 \sqrt{H^4 \left( \frac{H}{K \wedge H} \right) X_{i,h} A_i^{K\wedge H} T \log A_i}. \]
We first present the sample-based loss estimator for Type-I rehistories 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 $I \subset \mathbb{Z}_{\geq 1}$ and $n' \geq |I|$, $\text{fill}(I, n')$ is defined as the unique superset of $I$ with size $n'$ and the smallest possible additional elements, i.e.

$$\text{fill}(I, n') := I \cup \{I^c_1 \cup \ldots \cup I^c_{|I|}\},$$

(31)

where $I^c := \mathbb{Z}_{\geq 1} \setminus I$ with sorted elements $I^c_1 < I^c_2 < \ldots$. For example, $\text{fill}([1, 3, 8], 5) = \{1, 3, 8\} \cup \{2, 4\} = \{1, 2, 3, 4, 8\}$.

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 $\{S^t_b\}_{b \in \Omega^i_{x_{i,h}}}$ 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 $\text{fill}(\cdot, \cdot)$ defined in (31))

$$w_{b_{1:h-1}}(x_{i,h}) = \prod_{k \in \text{fill}(W \cup \{(h-1) \cup \{h\})} \pi_{i,k}^{t,h}(a_k | x_k), \quad \text{for all } b_{1:h-1} \in \Omega^i_{x_{i,h}};$$

(33)

$$w_{b_{1:h'}}(x_{i,h}) = \prod_{k \in W \cup \{(h+1, \ldots, h)\}} \pi_{i,k}^{t,h}(a_k | x_k), \quad \text{for all } b_{1:h'} \in \Omega^i_{x_{i,h}}.$$  

(34)

The resulting choice of time selection functions, $S^t_b = M^t_b w^t_b(x_{i,h})$, 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 $\pi^t_{i}$ 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 \( \pi_{t-1} \). Then the same procedure goes on for players 2, \ldots, 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) \((H+1)K^{H+1} + K \land H - 1 \leq 3H(K \land H)\) episodes.*

**Proof.** The proof follows by counting the number of \( W \)'s and \( W' \)'s in the sampling algorithms, i.e. the cardinalities of

\[
W_{1,h} := \{ W : W \subseteq [h-1] \text{ with } |W| = (K-1) \land (h-1) \}
\]

for \( 1 \leq h \leq H \), which comes from Line 1 in Algorithm 6, and

\[
W_{2,h',h} := \{ W : K \leq h' < h, W \subseteq [h'] \text{ with } |W| = K \text{ and ending in } h' \}
\]

for \( K \leq h \leq H \), which comes from line 1 in Algorithm 3.

For the first kind of sets, we have

\[
\sum_{h=1}^{H} |W_{1,h}| = \sum_{h=1}^{K \land H - 1} \left( \frac{h - 1}{(K - 1) \land (h - 1)} \right) + \sum_{h=K \land H}^{H} \left( \frac{h - 1}{(K - 1) \land (h - 1)} \right)
\]

\[
= K \land H - 1 + \sum_{h=K}^{H} \left( \frac{h - 1}{K \land H - 1} \right)
\]

\[
= K \land H - 1 + \left( \frac{H}{K \land H} \right).
\]

For the second kind of sets, we have

\[
\sum_{K \leq h' < h \leq H} |W_{2,h',h}| = \sum_{K \leq h' < h \leq H} \left( \frac{h' - 1}{K - 1} \right)
\]

\[
= \sum_{h=K}^{H} \sum_{h' = K}^{h-1} \left( \frac{h' - 1}{K - 1} \right) = \sum_{h=K}^{H} \left( \frac{h - 1}{K} \right)
\]

\[
= \left( \frac{H}{K + 1} \right) = \left( \frac{H}{K \land H + 1} \right).
\]

Taking summation gives that the number of episodes equals

\[
\sum_{h=1}^{H} |W_{1,h}| + \sum_{K \leq h' < h \leq H} |W_{2,h',h}| = \left( \frac{H + 1}{K \land H + 1} \right) + K \land H - 1.
\]

Finally, we show the above quantity can be upper bounded by \( 3H\left( \frac{H}{K \land H} \right) \). For \( K \geq H \), we have \( K \land H = H \) and the above quantity is \( 1 + H - 1 = H \leq 3H = 3H\left( \frac{H}{K \land H} \right) \). For \( K < H \), we have \( K \land H = K \), and thus

\[
\left( \frac{H + 1}{K \land H + 1} \right) + K \land H - 1 = \left( \frac{H + 1}{K + 1} \right) + K - 1 = \left( \frac{H}{K} \right) \cdot \left( \frac{H + 1}{K + 1} \right) + K - 1
\]

\[
\leq 2H \cdot \left( \frac{H}{K} \right) + H \leq 3H \cdot \left( \frac{H}{K} \right),
\]

where the last inequality follows from the fact that \( \left( \frac{H}{K} \right) \geq 1 \). This is the desired bound. \( \square \)
We bound the terms presented in Section G.4 & G.5 respectively. By the regret decomposition for conversion (Lemma E.1). The proof follows by bounding the regret (6):

\[ R_{i,K}^T = \max_{\phi \in \Phi^K_i} \sum_{t=1}^{T} \left( I_i^{\phi \pi_t^i \times \pi_{t-1}^i} - V_t^{\pi_t^i} \right) \]

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

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

\[ R_h^T := \max_{\phi \in \Phi^K_h} \sum_{x_h \in X_i,h} G_h^{T,\text{swap}}(x_h; \phi) + \max_{\phi \in \Phi^K_h} \sum_{x_h \in X_i,h} G_h^{T,\text{ext}}(x_h; \phi). \]

We bound the terms

\[ \sum_{h=1}^{H} \max_{\phi \in \Phi^K_h} \sum_{x_h \in X_i,h} G_h^{T,\text{swap}}(x_h; \phi) \text{ and } \sum_{h=1}^{H} \max_{\phi \in \Phi^K_h} \sum_{x_h \in X_i,h} G_h^{T,\text{ext}}(x_h; \phi) \]

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

\[ \eta_{x_h} = \sqrt{\left( \frac{H}{K \land H} \right) X_i A_i^{K \land H+1} t / (H^3 T)} \]

for all \( x_h \in X_i \) (same with (10)). With probability at least \( 1 - p/2 \), we have

\[ \sum_{h=1}^{H} \max_{\phi \in \Phi^K_h} \sum_{x_h \in X_i,h} G_h^{T,\text{swap}}(x_h; \phi) \leq \mathcal{O} \left( \sqrt{H^3 \left( \frac{H}{K \land H} \right) A_i^{K \land H+1} X_i T t} \right) + \mathcal{O} \left( \left( \frac{H}{K \land H} \right) A_i^{K \land H+1} X_i t \right) \]
with probability at least 

\[ \frac{1}{\epsilon^2} \]

gives

\[ R_t \]

Further using the "trivial" bound

\[ \log(8) \]

for all

\[ X \]

By Lemma G.2, G.3, and a union bound for all

\[ X \]

we get

\[ i \]

with bandit feedback)

\[ \sum_{h=1}^{H} \max_{\phi \in \Phi^k} G_{h,i}^{T,\text{ext}}(x_h;\phi) \leq O \left( \left( \frac{H^3}{K_i H} A_i^{K_i^H + 1} X_t \right) + O \left( \left( \frac{H^2}{H \wedge H} A_i^{K_i^H + 1} X_t \right) \right) \right. 

\]

\[ + O \left( \left( \frac{H^3}{K_i H} A_i^{K_i^H + 1} X_t \right) \right. 

\]

with probability at least \( 1 - \frac{p}{2} \) for all \( i \in [m] \) simultaneously, where \( \epsilon = \log(8 \sum_{j \in [m]} X_j A_j / p) \).

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

\[ R_{i,K}^T \leq HT \cdot \min \left\{ 1, O \left( \left( \frac{H^3}{K_i H} A_i^{K_i^H + 1} X_t \right) \right) \right\} 

\]

\[ \leq O \left( \left( \frac{H^3}{K_i H} A_i^{K_i^H + 1} X_t \right) \right. 

\]

where (i) follows by noticing that:

\[ T < H \left( \frac{H}{K \wedge H} X_i A_i^{K_i^H + 1} \right) \]

\[ R_{i,K}^T \leq HT \cdot \min \left\{ 1, O \left( \sqrt{H \left( \frac{H}{K \wedge H} X_i A_i^{K_i^H + 1} \right) / T} \right) \right\} 

\]

\[ T \geq H \left( \frac{H}{K \wedge H} X_i A_i^{K_i^H + 1} \right) \]

Therefore, as long as

\[ T \geq O \left( H^3 \left( \frac{H}{K \wedge H} \right) \left( \max_{i \in [m]} X_i A_i^{K_i^H + 1} \right) / \epsilon^2 \right) 

\]

40
we have by the online-to-batch lemma (Lemma E.1) that the average policy \( \pi = \text{Unif}\{\{\pi^t\}_{t=1}^T\} \) satisfies
\[
K\text{-EFCEGap}(\pi) = \frac{\max_{i \in [m]} R_{i,K}^T}{T} \leq \max_{i \in [m]} \mathcal{O}\left( H^3 \left( \frac{H}{K \wedge H} \right)^2 \left( \max_{i \in [m]} X_i A_i^{K \wedge H+1} \right) \right) \frac{1}{T} \leq \varepsilon.
\]
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 \( 3H(K \wedge H) \) episodes. Therefore, each iteration plays at most \( 3mH(K \wedge H) \) episodes, and so the total number of episodes played by Algorithm 7 is
\[
3mH(K \wedge H), T = \mathcal{O}\left( mH^4 \left( \frac{H}{K \wedge H} \right)^2 \left( \max_{i \in [m]} X_i A_i^{K \wedge H+1} \right) \right) \frac{1}{\varepsilon^2}.
\]
This is the desired result. \( \square \)

G.4 Proof of Lemma G.2

Proof. Recall that \( G_{T,\text{swap}}^T(x_h; \phi) \) is defined as (eq. (21))
\[
G_{T,\text{swap}}^T(x_h; \phi) := \sum_{b_{1:h-1}} \delta^{1-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \tilde{R}_{T,\text{swap}}^T(x_h, b_{1:h-1}).
\]
For each \( h \in [H] \) and \( (x_h, b_{1:h-1}) \in \Omega_1(K) \), we have
\[
\tilde{R}_{T,\text{swap}}^T(x_h, b_{1:h-1}) = \max_{\phi \in \Phi^K} \sum_{x_{1:h}, b_{1:h-1}} \delta^{1-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \lambda_{T,\text{swap}}(x_h, b_{1:h-1}).
\]
Substituting this into \( \max_{\phi \in \Phi^K} \sum_{x_h \in X_{1:h}} G_{T,\text{swap}}^T(x_h; \phi) \) yields that
\[
\max_{\phi \in \Phi^K} \sum_{x_h \in X_{1:h}} G_{T,\text{swap}}^T(x_h; \phi) = \max_{\phi \in \Phi^K} \sum_{x_{1:h}} \sum_{b_{1:h-1}} \delta^{1-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \lambda_{T,\text{swap}}(x_h, b_{1:h-1})
\]
\[
\leq \text{REGRET}_{T,\text{swap}} + \text{BIAS}_{T,\text{swap}} + \text{BIAS}_{2,T,\text{swap}},
\]
where
\[
\text{REGRET}_{T,\text{swap}} := \max_{\phi \in \Phi^K} \sum_{x_{1:h} \in X_{1:h}} \sum_{b_{1:h-1}} \delta^{1-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})
\]
\[
\times \max_{\phi \in \Phi^K} \sum_{x_{1:h}} \sum_{b_{1:h-1}} \delta^{1-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \lambda_{T,\text{swap}}(x_h, b_{1:h-1}).
\]
\[
\text{BIAS}_{T,\text{swap}} := \max_{\phi \in \Phi^K} \sum_{x_{1:h} \in X_{1:h}} \sum_{b_{1:h-1}} \delta^{1-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k})
\]
\[
\times \sum_{\phi \in \Phi^K} \sum_{x_{1:h}} \sum_{b_{1:h-1}} \delta^{1-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k}) \lambda_{T,\text{swap}}(x_h, b_{1:h-1}).
\]

[41]
The bounds for $\sum_{h=1}^{H} \regret_h$, $\sum_{h=1}^{H} \bias_{T,\text{swap}}^{1,h}$, and $\sum_{h=1}^{H} \bias_{T,\text{swap}}^{2,h}$ 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 $\regret_h$). If we choose learning rates as

$$\eta_{x_h} = \sqrt{\left( \frac{H}{K \wedge H} \right) X_i A_i^{K \wedge H+1} \log(X_i A_i/p)/(H^3T)},$$

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

$$\left| \sum_{h=1}^{H} \bias_{T,\text{swap}}^{h} \right| \leq \sqrt{H^3 \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H+1} X_i T T}$$

+ $O \left( \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H} X_i T T \right)$

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

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

$$\left| \sum_{h=1}^{H} \bias_{1,h}^{T,\text{swap}} \right| \leq O \left( \sqrt{H^3 \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H+1} X_i T T} \right) + H \left( \frac{H}{K \wedge H - 1} A_i^{K \wedge H} X_i T T \right),$$

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

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

$$\left| \sum_{h=1}^{H} \bias_{2,h}^{T,\text{swap}} \right| \leq O \left( \sqrt{H^3 \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H+1} X_i T T} \right) + H \left( \frac{H}{K \wedge H} A_i^{K \wedge H+1} X_i T T \right),$$

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

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

$$\left| \sum_{h=1}^{H} \bias_{T,\text{swap}}^{h}(x_h; \phi) \right| \leq \sum_{h=1}^{H} \regret_h \sum_{h=1}^{H} \bias_{1,h}^{T,\text{swap}} + \sum_{h=1}^{H} \bias_{2,h}^{T,\text{swap}}$$

\begin{align*}
&\leq O \left( \sqrt{H^3 \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H+1} X_i T T} \right) \\
&+ O \left( \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H} X_i T T \right) \sqrt{H \left( \frac{H}{K \wedge H} A_i^{K \wedge H+1} X_i T T \right)} \\
&+ O \left( \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H+1} X_i T T \right) .
\end{align*}
G.4.1 Proof of Lemma G.4: Bound on \( \text{REGRET}\_h^{T,\text{swap}} \)

Proof. Recall that \( \text{REGRET}\_h^{T,\text{swap}} \) is defined as

\[
\max_{\phi \in \Phi_h^{T,\text{swap}}} \sum_{x_h \in X_i} \sum_{b_l \in b_{1,h-1}} \delta^h \Delta_{l-1} (a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k}) 
\times \max_{\phi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_t^i (b_k | x_k) \left( \pi_t^i (\cdot | x_h) - \phi \circ \pi_t^i (\cdot | x_h) \right) \tilde{L}(x_h, b_{1,h-1}).
\]

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

\[
\max_{\phi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_t^i (b_k | x_k) \left( \pi_t^i (\cdot | x_h) - \phi \circ \pi_t^i (\cdot | x_h) \right) \tilde{L}(x_h, b_{1,h-1}).
\]

Recall that in Algorithm 7, each regret minimizer \( R_{x_h} \) is associated with weights \( M_{b_{1,h-1}} = \prod_{k=1}^{h-1} \pi_t^i (b_k | x_k) \) and \( w_{b_{1,h-1}} = \prod_{k \in \text{fill}(W, (h-1) \land (K-1))} \pi_t^i (a_k | x_k) \) for any \( b_{1,h-1} \in \Omega^{(1),K}_i (x_h) \), where \( W = \{ k \in [h-1] : b_k \neq a_k \} \). Let \( \mathcal{W} = \text{fill}(W, (h-1) \land (K-1)) \), \( \eta_{x_h} = \eta \) be the learning rate of \( R_{x_h} \) and \( L = H \). Since regret minimizers \( R_{x_h} \) observe \( \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) \) and \( \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) \) from Algorithm 3 & 6 as its loss vector at round \( t \), we have

\[
M_{b_{1,h-1}} w_{b_{1,h-1}} \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) = \frac{1}{H} \sum_{h' = h}^{H} \left( 1 - r_{t,h}^{i,\eta} (\mathcal{W}) \right)
\]

Similarly, we have \( M_{b_{1,h-1}} w_{b_{1,h-1}} \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) \in [0, L] \). Moreover, let \( \mathcal{F}_t \) be the \( \sigma \)-algebra containing all the information until \( \pi_t^i \) is sampled, by the sampling algorithm, we have that the loss estimators are unbiased:

\[
\mathbb{E} \left[ \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) | \mathcal{F}_{t-1} \right] = \mathbb{E} \left[ \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) | \mathcal{F}_{t-1} \right] = L^i_{t} (x_h, t),
\]

for all \( (x_h, b_{1,h-1}) \in \Omega^{(1),K}_i \) and all \( (x_h, b_{1,h'}) \in \Omega^{(1),K}_i \). So the assumptions in Lemma A.3 are satisfied. By Lemma A.3, with probability at least \( 1 - \frac{p}{8} \), for all \( x_h \in X_i \), we have

\[
\max_{\phi} \sum_{t=1}^{T} \prod_{k=1}^{h-1} \pi_t^i (b_k | x_k) \left( \pi_t^i (\cdot | x_h) - \phi \circ \pi_t^i (\cdot | x_h) \right) \tilde{L}(x_h, b_{1,h-1}) \leq \frac{2A_i \log (8X_i A_i / p)}{\eta w_{b_{1,h-1}}} + \eta \sum_{t=1}^{T} M_{b_{1,h-1}} \mathbb{E} \left[ \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) \right]
\]

\[
= \frac{2A_i \log (8X_i A_i / p)}{\eta \prod_{k \in \mathcal{W}_h} \pi_t^i (a_k | x_k)} + H \eta \prod_{k \in \mathcal{W}_h} \pi_t^i (b_k | x_k)
\times \sum_{t=1}^{T} \left( \pi_t^i (\cdot | x_h), \mathbb{E} \left[ \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) \right] \right)
\]

\[
\leq \frac{2A_i \log (8X_i A_i / p)}{\eta \prod_{k \in \mathcal{W}_h} \pi_t^i (a_k | x_k)} + H^2 \eta \sum_{t=1}^{T} \left( \pi_t^i (\cdot | x_h), \mathbb{E} \left[ \tilde{L}_{(x_h, b_{1,h-1})} (\cdot) \right] \right).
\]
Above, we used (i) our choices of \( M'_{b_1, \ldots, 1} \) and \( w_{b_1, \ldots, 1} \); (ii) \( (|B^s| + |B^c|)|\Psi^s| \leq A_{t+1, \ldots, t} \) and (iii) taking union bound over all infosets. Plugging this into \( \text{REGRET}^{T, \text{swap}}_h \), we have

\[
\text{REGRET}^{T, \text{swap}}_h \leq \max_{\phi \in \Phi^K} \sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \delta^{K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \\
\times \max_{\delta^{K-1}(a_{1:h-1}, b_{1:h-1})} \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \prod_{k \in [K \setminus h]} \pi^{*, h}_k(a_k | x_k) \\
\leq \frac{2A_1 \log(8X_tA_i/p)}{\eta} \max_{\phi \in \Phi^K} \sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \delta^{K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \\
\times \sum_{t=1}^{T} \prod_{k \in [K \setminus h]} \pi^{*, h}_k(b_k | x_k) \left( \pi^{*, h}_h(b_h | x_h), 1 \left\{ (x_h, \cdot) = (x^{t(h, \mathcal{W})}_h, a^{t(h, \mathcal{W})}_h) \right\} \right).
\]

Letting

\[
I_h := \frac{2A_1 \log(8X_tA_i/p)}{\eta} \max_{\phi \in \Phi^K} \sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \delta^{K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \prod_{k \in [K \setminus h]} \pi^{*, h}_k(a_k | x_k); \\
\Pi_h := H^2 \eta \max_{\phi \in \Phi^K} \sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \delta^{K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \sum_{t=1}^{T} \pi^{*, h}_h(a_k | x_k) \\
\times \sum_{k \in [K \setminus h]} \pi^{*, h}_k(b_k | x_k) \left( \pi^{*, h}_h(b_h | x_h), 1 \left\{ (x_h, \cdot) = (x^{t(h, \mathcal{W})}_h, a^{t(h, \mathcal{W})}_h) \right\} \right).
\]

Using Lemma B.4, we have

\[
\sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \delta^{K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \\
= \sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \delta^{K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \prod_{k \in \mathcal{W} \setminus [h]} \pi^{*, h}_k(a_k | x_k) \\
\sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \sum_{a_{h-1}} \delta^{K-1}(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k}) \prod_{k \in \mathcal{W} \setminus [h]} \pi^{*, h}_k(a_k | x_k) \\
= \sum_{x_h \in X_t, h_{b_1, h_{b_1}} - 1} \sum_{x_{h-1}} A_i^{[\mathcal{W} \setminus [h]]} \times \sum_{x_h, a_{h-1}} \left( \prod_{k \in \mathcal{W} \setminus [h]} \pi^{*, h}_k(a_k | x_k) \right)^{-1} \\
\times \sum_{b_{1:h-1}, \mu \in \mathcal{W} \setminus (K \setminus h) = \mathcal{W} \setminus [h]} \prod_{k \in \mathcal{W} \setminus [h]} \phi_k(a_k | x_k, b_{1:k}) \prod_{k \in \mathcal{W} \setminus [h]} \pi^{*, h}_k(b_k | x_k) \prod_{k \in \mathcal{W} \setminus [h]} \pi^{*, h}_k(a_k | x_k) \\
\sum_{x_{h-1}, \mu \in \mathcal{W} \setminus (K \setminus h)} A_i^{[\mathcal{W} \setminus [h]]} X_{t, h_{b_1, h_{b_1}}} A_i \\
\times X_{t, h_{b_1, h_{b_1}}} \left( \frac{h-1}{(K-1) \setminus (h-1)} \right) A_i^{K \setminus h}.
\]

Here, (i) uses that \( \pi^{*, h}_k (\cdot | x_k) \) is uniform distribution on \( A_k \) and that for \( k \in [h-1] \setminus \mathcal{W} \), we have \( a_k = b_k \); (ii) follows from grouping \( x_k \) and \( b_{1:k} \) by \( \mathcal{W} \); where \( \pi^{*, \text{unif}}_{k, h} \) is the uniform distribution on \( A_k \); (iii) uses Lemma B.4 and the fact that, for each fixed \( \mathcal{W} \) (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 $\pi^{*,h}_{i,k}(-|x_k)$ if step $k \in W$. Otherwise, sample recommended action $b_k$ from $\pi^{\text{unif}}_{i,k}(-|x_k)$.

• “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,

$$\sum_{x_k \in X_h} \sum_{b_{1:h-1}} \delta \leq K^{-1}(a_{1:h-1},b_{1:h-1}) \prod_{k \in W \cup \{h\}} \pi^{*,h}_{i,k}(a_k|x_k) \leq X_{i,h} \left( \frac{h-1}{K \wedge h-1} \right) A_i^{K \wedge h}. \quad (35)$$

So we have

$$I_h \leq \frac{2A_i \log(8X_iA_i/p)}{\eta} X_{i,h} \left( \frac{h-1}{(K-1) \wedge (h-1)} \right) A_i^{K \wedge h}.$$ 

For $\Pi_h$, observe that the random variables $\Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}}$ satisfy the following:

• $\Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}} = \prod_{k \in W} \pi_{i,k}^{x_{h},b_{1:k}}(b_k|x_k) \left( \pi^{t_{i,h}}(\cdot|x_k), 1 \left\{ (x_h, \cdot) = (x^{t_{i,h}}(h,W), a^{t_{i,h}}(h,W)) \right\} \right) \in [0, 1]$;

• Let $\mathcal{F}_{t-1}$ be the $\sigma$-algebra containing all information until $\pi^t$ is sampled, then

$$E \left( \Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}} \left| \mathcal{F}_{t-1} \right. \right) = \prod_{k \in W} \pi_{i,k}^{x_{h},b_{1:k}}(b_k|x_k) E \left[ \sum_{a \in A_i} \pi_i^t(a|x_k) 1 \left\{ (x_h, a) = (x^{t_{i,h}}(h,W), a^{t_{i,h}}(h,W)) \right\} \left| \mathcal{F}_{t-1} \right. \right]$$

$$= \prod_{k \in W} \pi_{i,k}^{x_{h},b_{1:k}}(b_k|x_k) \sum_{a \in A_i} \pi_i^t(a|x_k) \left( \prod_{k \in W} \pi_{i,k}^{x_{h},b_{1:k}}(a_k|x_k) \cdot \prod_{k \in [h-1]\setminus W} \pi_{i,h}^t(a_k|x_k) \cdot \pi_{i,k}^{x_{h},b_{1:k}}(a_k|x_k) \right)$$

$$= \prod_{k \in [h-1]} \pi_{i,k}^{x_{h},b_{1:k}}(b_k|x_k) \cdot \prod_{k \in W \cup \{h\}} \pi_{i,h}^t(a_k|x_k) \cdot \pi_{i,k}^{x_{h},b_{1:k}}(x_h),$$

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

• The conditional variance $E\left( (\Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}})^2 \left| \mathcal{F}_{t-1} \right. \right)$ can be bounded as

$$E \left[ (\Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}})^2 \left| \mathcal{F}_{t-1} \right. \right] \leq E \left[ \Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}} \left| \mathcal{F}_{t-1} \right. \right] = \prod_{k \in [h-1]} \pi_{i,k}^{x_{h},b_{1:k}}(b_k|x_k) \prod_{k \in W \cup \{h\}} \pi_{i,h}^t(a_k|x_k) \pi_{i,k}^{x_{h},b_{1:k}}(x_h),$$

where we used $\Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}} \in [0, 1]$ 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 $(h, x_{h}, b_{1:h-1})$:

$$\sum_{t=1}^T \Sigma_{\pi_{i,h}}^{x_{h},b_{1:h-1}} \leq (\lambda + 1) \sum_{t=1}^T \prod_{k \in W} \pi_{i,k}^{x_{h},b_{1:k}}(b_k|x_k) \prod_{k \in W \cup \{h\}} \pi_{i,h}^t(a_k|x_k) \pi_{i,k}^{x_{h},b_{1:k}}(x_h) + \frac{C \log(X_iA_i/p)}{\lambda},$$

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

$$\Pi_h \leq H^2 \eta \max_{\phi \in \Phi} \sum_{x_k \in X_{i,h}} \sum_{b_{1:h-1}} \delta \leq K^{-1}(a_{1:h-1},b_{1:h-1}) \prod_{k \in W \cup \{h\}} \pi_{i,k}^{x_{h},b_{1:h-1}}(a_k|x_k) \leq \prod_{k \in W \cup \{h\}} \pi_{i,k}^{x_{h},b_{1:h-1}}(a_k|x_k).$$
G.4.2 Proof of Lemma G.5: Bound on $h$ ≤ so we have $\lambda$

Then, choosing $\eta$

Taking summation over $i = 1$

so we have

max $\phi \in \Phi^K_i$

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

Using these two inequalities, we can get that

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

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

Then, choosing $\eta = \sqrt{(H_{K \wedge H}^1)}A_{i}^{K \wedge H + 1}X_i(t)/(H^3T)$ and using $(H^{-1}_{K \wedge H}) \leq (H_{K \wedge H}^1)$ we have

with probability at least $1 - p/8$, where $\epsilon = \log(8X_iA_i/p)$ is a log factor.

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

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

$$\text{BIAS}^{T,\text{swap}}_{1,h} = \max_{\phi \in \Phi^K_i} \sum_{x_h \in X_i,h} \sum_{b_1:h-1} \delta_{1:h-1}^{K-1} \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_1:k)}{\prod_{k \in [h-1]} \pi^{x_k}_{i,h}(a_k|x_k)}$$

$$\times \left[ (\lambda + 1) \sum_{t=1}^{T} \prod_{k \in [h-1]} \pi^{x_k}_{i,k}(b_k|x_k) \prod_{k \in [h-1]} \pi^{x_k}_{i,h}(a_k|x_k) \cdot p_{1,h}^{X_i}(x_i,h) + \frac{C \log(X_iA_i/p)}{\lambda} \right].$$

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

$$\sum_{x_h \in X_i,h} \sum_{b_1:h-1} \delta_{1:h-1}^{K-1} \frac{\prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_1:k)}{\prod_{k \in [h-1]} \pi^{x_k}_{i,k}(a_k|x_k)} \leq 1.$$
The conditional variance

By the definition of $\bar{L}$ in Algorithm 6, we can rewrite $\bar{\Delta}_t^{x_h,b_1,h-1}$ as

$$\bar{\Delta}_t^{x_h,b_1,h-1} = \prod_{k \in \mathcal{W} \cup \{h\}} \pi_{i,k}^t(a_k|x_k) \prod_{k=1}^{h-1} \pi_{i,k}^t(b_k|x_k) \left( \pi_i^t(\cdot|x_{i,h}), L_i^t(x_{i,h}, \cdot) - \bar{L}_i^t(x_{i,h}, b_{1:h-1}) \right).$$

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 $\overline{W} = \text{fill}(W; (h - 1) \land (K - 1))$. Observe that the random variables $\bar{\Delta}_t^{x_h,b_1,h-1}$ satisfy the following:

- By the definition of $\bar{L}$ in Algorithm 6, we can rewrite $\bar{\Delta}_t^{x_h,b_1,h-1}$ as

$$\bar{\Delta}_t^{x_h,b_1,h-1} = \prod_{k \in \mathcal{W} \cup \{h\}} \pi_{i,k}^t(a_k|x_k) \prod_{k=1}^{h-1} \pi_{i,k}^t(b_k|x_k) \left( \pi_i^t(\cdot|x_{i,h}), L_i^t(x_{i,h}, \cdot) - \bar{L}_i^t(x_{i,h}, b_{1:h-1}) \right)$$

- $\bar{\Delta}_t^{x_h,b_1,h-1} \leq \left( \pi_i^t(\cdot|x_{i,h}), F_{i,h}(x_{i,h}, \cdot) \right) \leq H$;

- $E[\bar{\Delta}_t^{x_h,b_1,h-1} | F_{t-1}] = 0$, where $F_{t-1}$ is the $\sigma$-algebra containing all information up to $t^t$ is sampled. This also can be seen from the unbiasedness of $\bar{L}$;

- The conditional variance $E[(\bar{\Delta}_t^{x_h,b_1,h-1})^2 | F_{t-1}]$ can be bounded as

$$E \left[ \left( \bar{\Delta}_t^{x_h,b_1,h-1} \right)^2 \bigg| F_{t-1} \right] \leq H^2 \prod_{k \in \mathcal{W}} \pi_{i,k}^t(b_k|x_k) \cdot E \left[ \left( \pi_i^t(\cdot|x_{i,h}), 1 \left\{ (x_{i,h}, \cdot) = (x_h^{i,h}, a_i^{i,h}) \right\} \right)^2 \bigg| F_{t-1} \right]$$

$$= H^2 \prod_{k \in \mathcal{W}} \pi_{i,k}^t(b_k|x_k) \cdot E \left[ \sum_{a \in A_i} \pi_i^t(a|x_k) 1 \left\{ (x_{i,h}, a) = (x_h^{i,h}, a_i^{i,h}) \right\} \bigg| F_{t-1} \right]$$

$$= H^2 \prod_{k \in \mathcal{W}} \pi_{i,k}^t(b_k|x_k) \cdot \sum_{a \in A_i} \pi_i^t(a|x_k) \prod_{k \in \mathcal{W} \cup \{h\}} \pi_i^t(a_k|x_k) \cdot \prod_{k \in [h-1] \setminus \mathcal{W}} \pi_i^t(a_k|x_k) \cdot \pi_i^t(a|x) \pi_i^t(a|x)$$

$$\leq H^2 \prod_{k \in [h-1] \setminus \mathcal{W}} \pi_{i,k}^t(b_k|x_k) \cdot \prod_{k \in \mathcal{W} \cup \{h\}} \pi_i^t(a|x_k).$$

Here, (i) uses $\left\{ \pi_i^t(\cdot|x_{i,h}), 1 \left\{ (x_{i,h}, \cdot) = (x_h^{i,h}, a_i^{i,h}) \right\} \right\} \in [0, 1]$; (ii) is because $\prod_{k \in \mathcal{W}} \pi_{i,k}^t(b_k|x_k) \cdot \prod_{k \in [h-1] \setminus \mathcal{W}} \pi_{i,k}^t(a_k|x_k) = \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k).$

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 $(h, x_{i,h}, b_{1:h-1})$:

$$\sum_{t=1}^{T} \bar{\Delta}_t^{x_h,b_1,h-1} \leq \lambda H^2 \prod_{k \in \mathcal{W} \cup \{h\}} \pi_{i,k}^t(a_k|x_k) \sum_{t=1}^{T} \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k) \pi_i^t(\cdot|x_{i,h}) + \frac{C \log(XA_1/p)}{\lambda},$$

47
where $C > 0$ is some absolute constant. Plugging this bound into $\text{BIAS}^{T,\text{swap}}_{1,h}$ yields that, for all $h \in [H]$,

$$\text{BIAS}^{T,\text{swap}}_{1,h} \leq \max_{\phi \in \Phi_i^K} \sum_{x_h \in X_i, h \in [1,t]} \sum_{b_1, b_{h-1}} \delta^{\leq K-1}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{k-1}) \prod_{k \in [1,h-1]} \pi_{i,k}^T(a_k|x_k)$$

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

$$\leq \lambda H^2 \max_{\phi \in \Phi_i^K} \sum_{x_h \in X_i, h \in [1,t]} \sum_{b_1, b_{h-1}} \delta^{\leq K-1}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{k-1}) \prod_{k \in [1,h-1]} \pi_{i,k}^T(b_k|x_k)p_{1,h}^{-1}(x_i,h) + \frac{C \log(X_iA_i/p)}{\lambda} \prod_{x_h \in X_i, h \in [1,t]} \sum_{b_1, b_{h-1}} \delta^{\leq K-1}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{k-1}) \prod_{k \in [1,h-1]} \pi_{i,k}^T(b_k|x_k)p_{1,h}^{-1}(x_i,h).$$

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

$$\sum_{x_h \in X_i, h \in [1,t]} \sum_{b_1, b_{h-1}} \delta^{\leq K-1}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{k-1}) \prod_{k \in [1,h-1]} \pi_{i,k}^T(b_k|x_k)p_{1,h}^{-1}(x_i,h) \leq 1.$$

So we have

$$\lambda H^2 \max_{\phi \in \Phi_i^K} \sum_{x_h \in X_i, h \in [1,t]} \sum_{b_1, b_{h-1}} \delta^{\leq K-1}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{k-1}) \sum_{t=1,k \in [h-1]} \pi_{i,k}^T(b_k|x_k)p_{1,h}^{-1}(x_i,h) \leq \lambda H^2 T.$$

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

$$\sum_{x_h \in X_i, h \in [1,t]} \sum_{b_1, b_{h-1}} \delta^{\leq K-1}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{k-1}) \sum_{t=1,k \in [h-1]} \pi_{i,k}^T(b_k|x_k)p_{1,h}^{-1}(x_i,h) \leq X_{i,h} \left( \frac{h-1}{K-1} \wedge (h-1) \right) A_i^{K_{\wedge h}}.$$

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

$$\text{BIAS}^{T,\text{swap}}_{1,h} \leq \lambda H^2 T + \frac{C \log(X_iA_i/p)}{\lambda} X_{i,h} \left( \frac{h-1}{K-1} \wedge (h-1) \right) A_i^{K_{\wedge h}}.$$

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

$$\sum_{h=1}^{H} \text{BIAS}^{T,\text{swap}}_{1,h} \leq \lambda H^3 T + \frac{C \log(X_iA_i/p)}{\lambda} \sum_{h=1}^{H} X_{i,h} \left( \frac{h-1}{K-1} \wedge (h-1) \right) A_i^{K_{\wedge h}}$$

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

$$\leq \lambda H^3 T + \frac{C \log(X_iA_i/p)}{\lambda} X_i \left( H \wedge K \wedge H \wedge H \right) A_i^{K_{\wedge H}}.$$

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

$$\sum_{t=1}^{T} \text{BIAS}^{T,\text{swap}}_{1,h} \leq O \left( H^3 (H \wedge K \wedge H) A_i^{K_{\wedge H}} X_i T_t + H \left( H \wedge K \wedge H \right) A_i^{K_{\wedge H}} X_i t \right),$$

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

**G.3 Proof of Lemma G.6: Bound on $\text{BIAS}^{T,\text{swap}}_{2,h}$**

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

$$\text{BIAS}^{T,\text{swap}}_{2,h} = \max_{\phi \in \Phi_i^K} \sum_{x_h \in X_i, h \in [1,t]} \sum_{b_1, b_{h-1}} \delta^{\leq K-1}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{k-1}).$$

48
and thus can be merged into a single max over \( \phi \).

By the definition of \( \phi \),

\[
\max_{k=1} \sum_{t=1}^{T_h-1} \phi_k(a_k|x_k, b_{t:k})
\]

and

\[
\max_{k=1} \sum_{t=1}^{T-h-1} \phi_k(a_k|x_k, b_{t:k})
\]

resulting in

\[
\max_{k=1} \sum_{t=1}^{T-h-1} \phi_k(a_k|x_k, b_{t:k})
\]

Thus, the inner max over \( \phi_h \) and the outer max over \( \phi_{1:h-1} \) are separable and can be merged into a single max over \( \phi_{1:h} \).

Observe that the random variables \( \Delta_{x,h}^{t} \) satisfy the following:

- By the definition of \( \tilde{L} \) in Algorithm 6, we can rewrite \( \Delta_{x,h}^{t} \) as

\[
\Delta_{x,h}^{t} = \prod_{k \in \mathcal{W} \cup \{h\}} \pi_{i,h}^{t}(b_{k}|x_{k}) \mathbf{1}\left\{ (x_{h}, a_{h}) = (x_{h}^{t}(h, \mathcal{W}), a_{h}^{t}(h, \mathcal{W})) \right\} \cdot \left( H - h + 1 - \sum_{h' \neq h} r_{t}(h, h') \right)
\]

- \( \Delta_{x,h}^{t} \in [-H, H] \).

- \( \mathbb{E}[\Delta_{x,h}^{t} | \mathcal{F}_{t-1}] = 0 \), where \( \mathcal{F}_{t} \) is the \( \sigma \)-algebra containing all information until \( \pi^t \) is sampled.

- The conditional variance \( \mathbb{E}[\Delta_{x,h}^{t}^{2} | \mathcal{F}_{t-1}] \) can be bounded as

\[
\mathbb{E}\left[ \Delta_{x,h}^{t,1:h-1,a_h}^{2} \right] | \mathcal{F}_{t-1} \leq \mathbb{E}\left[ \left( H - h + 1 - \sum_{h' \neq h} r_{t}(h, h') \right) \prod_{k \in \mathcal{W} \cup \{h\}} \pi_{i,h}^{t}(b_{k}|x_{k}) \mathbf{1}\left\{ (x_{h}, a_{h}) = (x_{h}^{t}(h, \mathcal{W}), a_{h}^{t}(h, \mathcal{W})) \right\} \right] | \mathcal{F}_{t-1}
\]

\[
\leq \mathbb{E}\left[ \left( H - h + 1 - \sum_{h' \neq h} r_{t}(h, h') \right) \prod_{k \in \mathcal{W} \cup \{h\}} \pi_{i,h}^{t}(b_{k}|x_{k}) \mathbf{1}\left\{ (x_{h}, a_{h}) = (x_{h}^{t}(h, \mathcal{W}), a_{h}^{t}(h, \mathcal{W})) \right\} \right] | \mathcal{F}_{t-1}
\]

\[
= H^2 \prod_{k \in \mathcal{W} \cup \{h\}} \pi_{i,h}^{t}(b_{k}|x_{k}) \prod_{k \in \mathcal{W}} \pi_{i,h}^{t}(a_{k}|x_{k}) \cdot \prod_{k \in [h-1]|\mathcal{W}} \pi_{i,h}^{t}(a_{k}|x_{k}) p_{i,h}^{t}(x_{h})
\]
\( H^2 \prod_{k \in [h]} \pi_{i,k}^t(b_k|x_k)p_{x,t}^{\pi_t}(x_{i,h}) \prod_{k \in W \cup (h)} \pi_{i,k}^h(a_k|x_k). \)

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

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 \( (h, x_{i,h}, b_{1:h}, a_{1:h}) \):

\[
\sum_{t=1}^T \xi_{i,h}^{a_{1:h},b_{1:h}} \leq \lambda H^2 \prod_{k \in W \cup (h)} \pi_{i,k}^t(a_k|x_k) \sum_{t=1}^T \prod_{k \in [h]} \pi_{i,k}^t(b_k|x_k)p_{x,t}^{\pi_t}(x_{i,h}) + \frac{C \log(X_i A_i/p)}{\lambda},
\]

where \( C > 0 \) is some absolute constant. Plugging this bound into \( \text{BIAS}^T_{2,h} \) yields that, for all \( h \in [H] \) and \( \phi \in \Phi^h \),

\[
\text{BIAS}^T_{2,h} \leq \max_{\phi \in \Phi^h} \sum_{x_{1:h},b_{1:h} \in \Omega} \sum_{\delta \leq K-1} \left( \prod_{k=1}^h \phi_k(a_k|x_k,b_{1:k}) \prod_{k \in W \cup (h)} \pi_{i,k}^t(a_k|x_k) \right) \times \lambda H^2 \prod_{k \in W \cup (h)} \pi_{i,k}^t(a_k|x_k) \sum_{t=1}^T \prod_{k \in [h]} \pi_{i,k}^t(b_k|x_k)p_{x,t}^{\pi_t}(x_{i,h}) + \frac{C \log(X_i A_i/p)}{\lambda}.
\]

By Corollary D.2, for fixed \( t \) and any \( \phi \in \Phi^h \),

\[
\sum_{x_{1:h},b_{1:h} \in \Omega} \sum_{\delta \leq K-1} \left( \prod_{k=1}^h \phi_k(a_k|x_k,b_{1:k}) \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k)p_{x,t}^{\pi_t}(x_{i,h}) \right) \leq 1.
\]

so we have

\[
\sum_{x_{1:h},b_{1:h} \in \Omega} \sum_{\delta \leq K-1} \left( \prod_{k=1}^h \phi_k(a_k|x_k,b_{1:k}) \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k)p_{x,t}^{\pi_t}(x_{i,h}) \right) = \frac{1}{\lambda} \sum_{x_{1:h},b_{1:h} \in \Omega} \sum_{\delta \leq K-1} \left( \prod_{k=1}^h \phi_k(a_k|x_k,b_{1:k}) \prod_{k \in [h-1]} \pi_{i,k}^t(b_k|x_k)p_{x,t}^{\pi_t}(x_{i,h}) \right) \leq T.
\]

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

\[
\sum_{x_{1:h},b_{1:h} \in \Omega} \sum_{\delta \leq K-1} \left( \prod_{k=1}^h \phi_k(a_k|x_k,b_{1:k}) \prod_{k \in W \cup (h)} \pi_{i,k}^t(a_k|x_k) \right) \leq X_{i,h} \left( \frac{h - 1}{K \land H - 1} \right) A^h_{K \land h+1}.
\]

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

\[
\text{BIAS}^T_{2,h} \leq \lambda H^2 T + \frac{C \log(X_i A_i/p)}{\lambda} X_{i,h} \left( \frac{h - 1}{K \land H - 1} \right) A^h_{K \land h+1}.
\]

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

\[
\sum_{h=1}^H \text{BIAS}^T_{2,h} \leq \lambda H^3 T + \frac{C \log(X_i A_i/p)}{\lambda} \sum_{h=1}^H X_{i,h} \left( \frac{h - 1}{(K - 1) \land (h - 1)} \right) A^h_{K \land h+1}.
\]

50
For each maximal element in $\Phi$.

**Proof.** G.5 Proof of Lemma G.3

\[ \lambda H^3 + \frac{C \log(\lambda X_1 \lambda)}{\lambda} X_1 \left( \frac{H - 1}{K \wedge H - 1} \right) A_i^{K \wedge H + 1} \]

for all $\lambda \in (0, 1/H)$. Choose $\lambda = \min \left\{ \frac{1}{H}, \sqrt{\frac{C X_1 \lambda}{H / T}} \right\}$, we obtain the bound

\[ \sum_{t=1}^T \text{BIAS}_{2,h}^{T,\text{swap}} \leq O \left( H^3 \left( \frac{H}{K \wedge H} \right) A_i^{K \wedge H + 1} X_t \right), \]

where $t = \log(8X_iA_1/p)$ is a log factor. \hfill \square

**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 $G_{h}^{T,\text{ext}}(x_h; \phi)$ (eq. (22)) is defined as

\[ G_{h}^{T,\text{ext}}(x_h; \phi) := \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k \wedge \tau_K}) \tilde{R}_{(x_h,b_{1:h-1},x_h)}. \]

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

\[ \tilde{R}_{(x_h,b_{1:h})}^{T,\text{ext}} = \max_a \sum_{t=1}^T \prod_{k=1}^{h'} \pi^t_i(b_k|x_k) \left( \left( \pi^t_i, (\cdot | x_i,h) \right), L^t_i, h(x_i,h) \right) - L^t_i, h(x_i,h) \]

Substituting this into $\max_{\phi \in \Phi^K} \sum_{x_h \in X_i,h} G_{h}^{T,\text{ext}}(x_h; \phi)$ yields that

\[ \max_{\phi \in \Phi^K} \sum_{x_h \in X_i,h} G_{h}^{T,\text{ext}}(x_h; \phi) = \max_{\phi \in \Phi^K} \sum_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k \wedge \tau_K}) \tilde{R}_{(x_h,b_{1:h-1},x_h)}^{T,\text{ext}} \]

where

\[ \text{REGRET}_{h}^{T,\text{ext}} := \max_{\phi \in \Phi^K} \sum_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k \wedge \tau_K}) \]

\[ \times \max_a \sum_{t=1}^T \prod_{k=1}^{h''} \pi^t_i(b_k|x_k) \left( \left( \pi^t_i, (\cdot | x_i,h) \right), L^t_i, h(x_i,h) \right) - L^t_i, h(x_i,h) \]

\[ \text{BIAS}_{1,h}^{T,\text{ext}} := \max_{\phi \in \Phi^K} \sum_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k \wedge \tau_K}) \]

\[ \times \sum_{t=1}^T \prod_{k=1}^{h} \pi^t_i(b_k|x_k) \left( \left( \pi^t_i, (\cdot | x_i,h) \right), L^t_i, h(x_i,h) \right) - L^t_i, h(x_i,h) \]

\[ \text{BIAS}_{2,h}^{T,\text{ext}} := \max_{\phi \in \Phi^K} \sum_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k \wedge \tau_K}) \]
The bounds for $\sum_{h=1}^{H} \text{REGRET}_{h}^{T,\text{ext}}$, $\sum_{h=1}^{H} \text{BIAS}_{1,h}^{T,\text{ext}}$, and $\sum_{h=1}^{H} \text{BIAS}_{2,h}^{T,\text{ext}}$ 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 $\text{REGRET}_{h}^{T,\text{ext}}$). If we choose learning rates as

$$\eta_{x_{h}} = \sqrt{\left(\frac{H}{K \wedge H}\right) x_{i} A_{i}^{K \wedge H+1} \log(8 X_{i} A_{i}/p)/(H^{3} T)}$$

for all $x_{h} \in X_{i}$ (same with (10)). Then with probability at least $1 - p/4$, we have

$$\sum_{h=1}^{H} \text{REGRET}_{h}^{T,\text{ext}} \leq \sqrt{H^{3} \left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}} + O\left(\left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}\right).$$

where $\iota = \log(8 X_{i} A_{i}/p)$ is a log factor.

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

$$\sum_{h=1}^{H} \text{BIAS}_{1,h}^{T,\text{ext}} \leq O\left(\sqrt{H^{3} \left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}} + \left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}\right),$$

where $\iota = \log(8 X_{i} A_{i}/p)$ is a log factor.

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

$$\sum_{h=1}^{H} \text{BIAS}_{2,h}^{T,\text{ext}} \leq O\left(\sqrt{H^{3} \left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}} + \left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}\right),$$

where $\iota = \log(8 X_{i} A_{i}/p)$ is a log factor.

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

$$\sum_{h=1}^{H} \max_{\phi \in \Phi^{K}} \sum_{x_{h} \in X_{i}, h \in X_{i}} C_{h}^{T,\text{ext}}(x_{h}; \phi) \leq \sum_{h=1}^{H} \text{REGRET}_{h}^{T,\text{ext}} + \sum_{h=1}^{H} \text{BIAS}_{1,h}^{T,\text{ext}} + \sum_{h=1}^{H} \text{BIAS}_{2,h}^{T,\text{ext}}$$

$$\leq \sqrt{H^{3} \left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}} + O\left(\left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}\right) + O\left(\left(\frac{H}{K \wedge H}\right) A_{i}^{K \wedge H+1} X_{i} T_{i}\right).$$

**G.5.1 Proof of Lemma G.7: Bound on $\text{REGRET}_{h}^{T,\text{ext}}$**

**Proof.** Recall that $\text{REGRET}_{h}^{T,\text{ext}}$ is defined as

$$\max_{\phi \in \Phi^{K}} \sum_{x_{h} \in X_{i}, h \in X_{i}} \sum_{b_{1,h-1}} \delta_{h,h-1}^{K}(a_{1,h-1}, b_{1,h-1}) \prod_{k=1}^{h-1} \phi_{k}(a_{k}, b_{k} \wedge \tau_{K})$$

$$\times \max_{a} \sum_{k=1}^{T} \prod_{l=1}^{T} \pi_{l}^{k}(b_{k} | x_{k}) \left(\tilde{L}_{(\tilde{x}_{l}, b_{l} \wedge \tau_{K})}(a) - L_{(x_{l}, h, a)}^{k}(x_{l}, h, a)\right).$$
We first apply regret minimization lemma (Lemma A.3) on $\mathcal{R}_{x_h}$ to give an upper bound of
\[
\max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_h} \pi_t^i(b_k|x_k) \left( \left< \pi_t^i(\cdot|x_i,h), \tilde{L}_{\kappa}^i(x_h,b_{1:h},\tau_{\kappa}) \right> - \tilde{L}_{\kappa}(x_h,b_{1:h},\tau_{\kappa})(a) \right).
\]
Here, we use $\eta
\]
For $\mathcal{R}_{x_h}$, Algorithm 7 gives that $M_{b_{1:h'}}^i = \prod_{k=1}^{k'} \pi_t^i(b_k|x_k)$ and $w_{b_{1:h'}} = \prod_{k \in W \cup \{h'+1, \ldots, h\}} \pi_t^i(a_k|x_k)$ for any $b_{1:h'} \in \Omega_{\kappa}(x_h)$, where $W = \{k \in [h-1] : b_k \neq a_k\}$. Letting $W(h) = W \cup \{h'+1, \ldots, h\}$, $\eta$ be the learning rate of $\mathcal{R}_{x_h}$ and $\mathcal{T} = H$. 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 $x_h \in cX_i$,
\[
\max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_h} \pi_t^i(b_k|x_k) \left( \left< \pi_t^i(\cdot|x_i,h), \tilde{L}_{\kappa}^i(x_h,b_{1:h},\tau_{\kappa}) \right> - \tilde{L}_{\kappa}(x_h,b_{1:h},\tau_{\kappa})(a) \right) \leq \frac{2 \log(8X_iA_i/p)}{\eta \max_{b_{1:h'}} M_{b_{1:h'}}^i} + \eta \sum_{t=1}^{T} M_{b_{1:h'}}^i \mathcal{T} \left< \pi_t^i(\cdot|x_k), \tilde{L}_{\kappa}(x_h,b_{1:h},\tau_{\kappa}) \right> \right>
\]
\[
= \frac{2 \log(8X_iA_i/p)}{\eta \prod_{k \in W(h)} \pi_t^i(a_k|x_k)} + H \eta \prod_{k=1}^{\tau_h} \pi_t^i(b_k|x_k) \times \sum_{t=1}^{T} \left< \pi_t^i(\cdot|x_i,h), \frac{1}{\prod_{k \in W(h)} \pi_t^i(a_k|x_k)} \pi_t^i(\cdot|x_k) \right> \prod_{k \in W(h)} \pi_t^i(b_k|x_k) \left\{ (x_h, \cdot) = (x_{h},(h',h'),x_{h},(h',W)) \right\} \right>
\]
\[
\leq \frac{2 \log(8X_iA_i/p)}{\eta \prod_{k \in W(h)} \pi_t^i(a_k|x_k)} + H^2 \eta \prod_{t=1}^{T} \left< \pi_t^i(\cdot|x_k), \frac{1}{\prod_{k \in W(h)} \pi_t^i(a_k|x_k)} \pi_t^i(\cdot|x_k) \right> \prod_{k \in W(h)} \pi_t^i(b_k|x_k) \left\{ (x_h, \cdot) = (x_{h},(h',h'),x_{h},(h',W)) \right\} \right>. \]
Here, we use (i) our choices of $M_{b_{1:h'}}^i$ and $w_{b_{1:h'}}$, (ii) $|B^e| + |B^c||\Psi^e| \leq a_i + 1 \leq X_iA_i$ and (iii) taking union bound over all infosets. Plugging this into $\text{REGRET}^i$, we have
\[
\text{REGRET}^i = \max_{\phi \in \Phi^i} \sum_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k,h'}) \times \max_a \sum_{t=1}^{T} \prod_{k=1}^{\tau_h} \pi_t^i(b_k|x_k) \left( \left< \pi_t^i(\cdot|x_i,h), \tilde{L}_{\kappa}^i(x_h,b_{1:h},\tau_{\kappa}) \right> - \tilde{L}_{\kappa}(x_h,b_{1:h},\tau_{\kappa})(a) \right) \leq \frac{2 \log(8X_iA_i/p)}{\eta \prod_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k,h'})} \prod_{k \in W(h)} \pi_t^i(a_k|x_k) \times \sum_{t=1}^{T} \prod_{k \in W(h)} \pi_t^i(b_k|x_k) \left< \pi_t^i(\cdot|x_i,h), 1 \right> \left\{ (x_h, \cdot) = (x_{h},(h',h'),x_{h},(h',W)) \right\} \right>. \]
\[
= \frac{2 \log(8X_iA_i/p)}{\eta \prod_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k,h'})} \prod_{k \in W(h)} \pi_t^i(a_k|x_k) \times \sum_{t=1}^{T} \prod_{k \in W(h)} \pi_t^i(b_k|x_k) \left< \pi_t^i(\cdot|x_i,h), 1 \right> \left\{ (x_h, \cdot) = (x_{h},(h',h'),x_{h},(h',W)) \right\} \right>. \]
Letting
\[
I_h := \frac{2 \log(8X_iA_i/p)}{\eta \prod_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k,h'})} \prod_{k \in W(h)} \pi_t^i(a_k|x_k) \sum_{t=1}^{T} \sum_{k \in W(h)} \pi_t^i(a_k|x_k) \left< \pi_t^i(\cdot|x_i,h), 1 \right> \left\{ (x_h, \cdot) = (x_{h},(h',h'),x_{h},(h',W)) \right\} \right>. \]
\[
II_h := H^2 \eta \prod_{x_h \in X_i,h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1},b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k,b_{1:k,h'}) \prod_{k \in W(h)} \pi_t^i(a_k|x_k) \sum_{t=1}^{T} \sum_{k \in W(h)} \pi_t^i(a_k|x_k) \left< \pi_t^i(\cdot|x_i,h), 1 \right> \left\{ (x_h, \cdot) = (x_{h},(h',h'),x_{h},(h',W)) \right\} \right>. \]

53
Using Lemma B.4, we have

$$\sum_{x_h \in X_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1,h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k:h}) \prod_{k \in W(h)} \pi^*_i(k)(a_k | x_k)$$

$$= \sum_{x_h \in X_{i,h}} \sum_{b_{1:h-1}} \delta^K(a_{1,h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k | x_k, b_{1:k:h}) \prod_{k \in [h-1] \setminus W} \pi^*_i(k)(a_k | x_k)\prod_{k \in W} \pi^*_i(k)(a_k | x_k)$$

$$\leq \sum_{W \subseteq [h-1] \setminus |W|=K} \sum_{x_h, a_h} A_i^{W | X_{i,h}} A_i \left( \prod_{k \in |W|} \pi^*_i(k)(a_k | x_k) \right)^{-1}$$

$$\times \sum_{b_{1:h-1} : \{k \in [h-1] : a_k \neq b_k\} = W} \prod_{h=1}^{h-1} \phi_k(a_k | x_k, b_{1:k:h}) \prod_{k \in [h-1] \setminus W} \pi^*_i(k)(b_k | x_k) \prod_{k \in W} \pi^{\text{unif}}(b_k) \pi^*_i(k)(a_k | x_k)$$

$$\leq \sum_{W \subseteq [h-1] \setminus |W|=K} A_i^{W | X_{i,h}} A_i \left( \frac{h-1}{K} \right)^{A_i^K \wedge H+1}.$$
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 for all \((h, x_{i,h}, b_{1:h-1})\):

\[
\sum_{t=1}^{T} \Delta_t^{(x_{i,h}, b_{1:h-1})} \leq (\lambda + 1) \sum_{t=1}^{T} \prod_{k \in [h']} \pi_{i,k}(b_k | x_k) \prod_{k \in W(h)} \pi_{i,k}^{*h}(a_k | x_k) p_{1:h}^{*i}(x_h) + \frac{C \log(X_i A_i/p)}{\lambda},
\]

where \( C > 0 \) is some absolute constant. Plugging this bound into \( \Pi_h \) yields that,

\[
\Pi_h \leq H^2 \eta \max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}, b_{1:h-1}} \sum_{a_k} \delta^K(a_{i,h-1}, b_{1:h-1}) \prod_{k \in W(h)} \phi_k(a_k | x_k, b_{1:k}, \lambda) \prod_{k \in W(h)} \pi_{i,k}^{*h}(a_k | x_k) \prod_{k \in [h']} \pi_{i,k}(b_k | x_k) p_{1:h}^{*i}(x_h) \leq 1.
\]

So we have

\[
\max_{\phi \in \Phi^K} \sum_{x_h \in X_{i,h}, b_{1:h-1}} \sum_{a_k} \delta^K(a_{i,h-1}, b_{1:h-1}) \prod_{k \in W(h)} \phi_k(a_k | x_k, b_{1:k}, \lambda) \prod_{k \in [h']} \pi_{i,k}(b_k | x_k) p_{1:h}^{*i}(x_h) \leq T.
\]

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

\[
\sum_{x_h \in X_{i,h}, b_{1:h-1}} \sum_{a_k} \delta^K(a_{i,h-1}, b_{1:h-1}) \prod_{k \in W(h)} \phi_k(a_k | x_k, b_{1:k}, \lambda) \prod_{k \in [h']} \pi_{i,k}(b_k | x_k) p_{1:h}^{*i}(x_h) \leq X_{i,h} \left( h - 1 \right) A_i^{K \wedge H+1}.
\]

Using these two inequalities, we can get that

\[
\Pi_h \leq H^2 \eta (\lambda + 1) T + \frac{CH^2 \eta \log(X_i A_i/p)}{\lambda} X_{i,h} \left( h - 1 \right) A_i^{K \wedge H+1}.
\]

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

\[
\sum_{h=1}^{H} \text{REGRET}_{h} = \sum_{h=1}^{H} (I_h + \Pi_h) \leq H^3 \eta (\lambda + 1) T + \left( \frac{2 \log(8X_i A_i/p)}{\eta} + \frac{CH^2 \eta \log(X_i A_i/p)}{\lambda} \right) \sum_{h=1}^{H} X_{i,h} \left( h - 1 \right) A_i^{K \wedge H+1}.
\]
\[
\leq H^3 \eta (\lambda + 1)T + \left( \frac{2 \log(8X_tA_t/p)}{\eta} + \frac{CH^2 \eta \log(X_tA_t/p)}{\lambda} \right) X_t \left( \frac{H - 1}{K \wedge H} \right) A_t^{K \wedge H + 1}
\]
\[
\leq H^3 \eta (\lambda + 1)T + \left( \frac{2 \log(8X_tA_t/p)}{\eta} + \frac{CH^2 \eta \log(X_tA_t/p)}{\lambda} \right) X_t \left( \frac{H}{K \wedge H} \right) A_t^{K \wedge H + 1},
\]

for all \( \lambda \in (0, 1) \). Choosing \( \lambda = 1 \), we have,
\[
\sum_{h=1}^{H} \text{REGERT}_h^{T, \text{ext}} \leq 2H^3 \eta T + \left( \frac{2 \log(8X_tA_t/p)}{\eta} + \frac{CH^2 \eta \log(X_tA_t/p)}{\lambda} \right) X_t \left( \frac{H - 1}{K \wedge H} \right) A_t^{K \wedge H + 1}.
\]

Then, choosing \( \eta = \sqrt{\left( \frac{H}{K \wedge H} \right) A_t^{K \wedge H + 1} X_t / (H^3 T)} \), we have
\[
\sum_{h=1}^{H} \text{REGERT}_h^{T, \text{ext}} \leq \left[ H^3 \left( \frac{H}{K \wedge H} \right) A_t^{K \wedge H + 1} X_t T \right] + O \left( \left[ H \left( \frac{H}{K \wedge H} \right) A_t^{K \wedge H + 1} X_t / T \right] \right),
\]

where \( t = \log(8X_tA_t/p) \) is a log factor. \( \square \)

G.5.2 Proof of Lemma G.8: Bound on \( \text{BIAS}_t^{T, \text{ext}} \)

**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 \( W \cup \{ h' + 1, \ldots, h \} \). We can rewrite \( \text{BIAS}_t^{T, \text{ext}} \) as

\[
\text{BIAS}_t^{T, \text{ext}} = \max_{\phi \in \Phi^t} \sum_{x_h \in A_{1,h-1}} \sum_{b_{1, h-1}} \delta^K (a_{1, h-1}, b_{1, h-1}) \prod_{k=1}^{h-1} \phi_k (a_k | x_k, b_{1:k \wedge h'}) \prod_{k \in W(h)} \pi_{i,k}^t (a_k | x_k)
\]

\[
\times \prod_{t=1}^{T} \prod_{k \in W(h)} \pi_{i,k}^t (a_k | x_k) \sum_{k=1}^{h'} \pi_{i,k}^t (a_k | x_k) \left( \pi_{i,k}^t (\cdot | x_k), L_{i,k}^t (x_i, h', \cdot) - \tilde{L}_{i}^t (x_h, b_{1,h'}) \right)
\]

\[
:= \Delta_t^{(x_h, b_{1,h'})}
\]

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

- By the definition of \( \tilde{L} \), we can rewrite \( \Delta_t^{(x_h, b_{1,h'})} \) as
  \[
  \Delta_t^{(x_h, b_{1,h'})} = \prod_{k=1}^{h'} \pi_{i,k}^t (b_k | x_k) \left( \pi_{i,k}^t (\cdot | x_k), L_{i,k}^t (x_i, h', \cdot) \right) \prod_{k \in W(h)} \pi_{i,k}^t (a_k | x_k)
  \]
  \[
  - \prod_{k \in W} \pi_{i,k}^t (b_k | x_k) \left( \pi_{i,k}^t (\cdot | x_k), 1 \left\{ (x_{h'}, \cdot) = (x_{h'}^{(h', h', W)}, a_{h'}^{(h', h', W)}) \right\} \right) \left( H - h + 1 - \sum_{k' \in W} \pi_{i,k'}^t (h, h', W) \right)
  \]
- \( \Delta_t^{(x_h, b_{1,h'})} \leq \left( \pi_{i,k}^t (\cdot | x_k), L_{i,k}^t (x_h, b_{1,h'}) \right) \leq H ;
- \( E[\Delta_t^{(x_h, b_{1,h'})} | \mathcal{F}_{t-1}] = 0 \), where \( \mathcal{F}_{t-1} \) is the \( \sigma \)-algebra containing all information until \( \pi^t \) is sampled. This also can be seen from the unbiasedness of \( \tilde{L} ;
- The conditional variance \( E[(\Delta_t^{(x_h, b_{1,h'})})^2 | \mathcal{F}_{t-1}] \) can be bounded as
  \[
  E \left[ \Delta_t^{(x_h, b_{1,h'})} \right]^2 | \mathcal{F}_{t-1}
  \]

\[
56
\]
where with probability at least $\lambda H$

Note that for fixed $h \in \{h \mid \leq \lambda H \}$

Here, (i) uses $H^2 \prod_{k \in W} \pi_{t,k}(b_k|x_k) \mathbb{E}\left[\sum_{a \in A_i} \pi_{t,h}(a|x_h, 1 \{ (x_h, a) = (x_t^{t,h,h',W}, a_t^{t,h,h',W}) \}) \right] | F_{t-1}$

$= H^2 \prod_{k \in W} \pi_{t,k}(b_k|x_k) \mathbb{E}\left[\sum_{a \in A_i} \pi_{t,h}(a|x_h) \prod_{k \in W(h)} \pi_{t,h}(a_k|x_k) \cdot \prod_{k \in \{h'\} \setminus W} \pi_{t,h}(a_k|x_k) \cdot p_{1,h}(x_i,h) \right]$

$\leq H^2 \prod_{k \in \{h'\}} \pi_{t,k}(b_k|x_k) \prod_{k \in W(h)} \pi_{t,h}(a_k|x_k) = \prod_{k \in \{h'\}} \pi_{t,k}(b_k|x_k)$.

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 $(h, x_i,h, b_i, b_{i-1})$:

$$\sum_{t=1}^{T} \Delta(x_t, b_{t-1}, h, x_{h+1}) \leq \lambda H^2 \prod_{k \in W(h)} \pi_{t,h}(a_k|x_k) \sum_{t=1}^{T} \prod_{k \in \{h'\}} \pi_{t,k}(b_k|x_k) p_{1,h}(x_i,h) + C \log(X_i A_i/p)$$

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

$$\text{BIAS}_{1,h}^{T,\text{ext}} \leq \max_{\phi \in \Phi^K} \sum_{x_h \in X_h, b_{h-1}} \sum_{b_{h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k \in W(h)} ^{h-1} \phi_k(a_k|x_k) \prod_{k \in \{h'\}} \pi_{t,k}(b_k|x_k) p_{1,h}(x_i,h) + C \log(X_i A_i/p) \lambda$$

So we have

$$\lambda H^2 \max_{\phi \in \Phi^K} \sum_{x_h \in X_h, b_{h-1}} \sum_{b_{h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k \in W(h)} ^{h-1} \phi_k(a_k|x_k) \prod_{k \in \{h'\}} \pi_{t,k}(b_k|x_k) p_{1,h}(x_i,h) \leq \lambda H^2 T.$$

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

$$\sum_{x_h \in X_h, b_{h-1}} \sum_{b_{h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k \in W(h)} ^{h-1} \phi_k(a_k|x_k) \prod_{k \in \{h'\}} \pi_{t,k}(b_k|x_k) p_{1,h}(x_i,h) \leq \sum_{i=1}^{H} \binom{h-1}{K} A_{i}^{K \times h+1}.$$
Plugging these bounds into $\text{BIAS}^{T, \text{ext}}_{1, h}$ yields that, with probability at least $1 - p/8$, for all $h \in [H]$,

$$\text{BIAS}^{T, \text{ext}}_{1, h} \leq \lambda H^2 T + \frac{C \log(X_i A_i / p)}{\lambda} X_{i, k} \left( \frac{h - 1}{K} \right) A_i^{K \land H + 1}.$$

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

$$\sum_{h=1}^{H} \text{BIAS}^{T, \text{ext}}_{1, h} \leq \lambda H^3 T + \frac{C \log(X_i A_i / p)}{\lambda} \sum_{h=1}^{H} X_{i, h} \left( \frac{h - 1}{K} \right) A_i^{K \land H + 1},$$

for all $\lambda \in (0, 1/H]$. Choose $\lambda = \min \left\{ \frac{1}{H}, \sqrt{\frac{C X_i A_i / p}{H T}} \right\}$, we obtain

$$\sum_{t=1}^{T} \text{BIAS}^{T, \text{ext}}_{1, h} \leq O \left( H^3 \left( \frac{H}{K \land H} \right) A_i^{K \land H + 1} X_i T + H \left( \frac{H}{K \land H} \right) A_i^{K \land H + 1} X_i T \right),$$

where $t = \log(8X_i A_i / p)$ is a log factor.

\[\square\]

G.5.3 Proof of Lemma G.9: Bound on $\text{BIAS}^{T, \text{ext}}_{2, h}$

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 $W \cup \{h' + 1, \ldots, h\}$. We can rewrite $\text{BIAS}^{T, \text{ext}}_{2, h}$ as

$$\text{BIAS}^{T, \text{ext}}_{2, h} = \max_{\phi \in \Phi^h} \sum_{x_h \in X_h} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h-1} \phi_k(a_k|x_k, b_{1:k\land h'})$$

\[x = \max_{a} \sum_{t=1}^{T} \prod_{k=1}^{h'} \pi'_t(b_k|x_k) \left( \overline{L}_{(x_h, b_{1:h})}^{\tau_{x_h, b_{1:h'}}}(a_{h'}) - L_{(x_h, b_{1:h'})}^{\tau_{x_h, b_{1:h'}}}(a_{h'}) \right) \]

\[x = \max_{a} \sum_{a_{h'}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi(a_k|x_k, b_{1:k\land h'}) \]

\[x = \max_{a} \sum_{t=1}^{T} \prod_{k=1}^{h'} \pi'_t(b_k|x_k) \left( \overline{L}_{(x_h, b_{1:h'})}^{\tau_{x_h, b_{1:h'}}}(a_{h'}) - L_{(x_h, b_{1:h'})}^{\tau_{x_h, b_{1:h'}}}(a_{h'}) \right) \]

Here, (i) comes from the fact that the inner max over $\phi_h'$ 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 $\overline{\Delta}_{t}^h$ satisfy the following:
• By the definition of \( \bar{L} \), we can rewrite \( \Delta_{t}^{\pi_{t}, b_{1:t}, \cdot, \cdot, \cdot} \) as

\[
\Delta_{t}^{\pi_{t}, b_{1:t}, \cdot, \cdot, \cdot} = \prod_{k \in W} \pi_{i,k}^{t}(b_{k}|x_{k}) \left\{ (x_{k}, a_{h}) = (x_{k}^{t,(h, \cdot, \cdot, \cdot)}, a_{h}^{t,(h, \cdot, \cdot, \cdot)}) \right\} \left( H - h + 1 - \sum_{t' = h}^{H} \pi_{i,t'}^{t,(h, \cdot, \cdot, \cdot)} \right) - \prod_{k \in [h']} \pi_{i,k}^{t}(b_{k}|x_{k}) L_{t}(x_{k}, b_{1:h'}) (a_{h}) \prod_{k \in W(h)} \pi_{i,k}^{t,h}(a_{k}|x_{k}).
\]

• \( \Delta_{t}^{\pi_{t}, b_{1:t}, \cdot, \cdot, \cdot} \leq H. \)

• \( \mathbb{E}[\Delta_{t}^{\pi_{t}, b_{1:t}, \cdot, \cdot, \cdot} | \mathcal{F}_{t-1}] = 0 \), where \( \mathcal{F}_{t-1} \) is the \( \sigma \)-algebra containing all information until \( \pi_{t} \) is sampled.

• The conditional variance \( \mathbb{E}[(\Delta_{t}^{\pi_{t}, b_{1:t}, \cdot, \cdot, \cdot})^{2} | \mathcal{F}_{t-1}] \) can be bounded as

\[
\mathbb{E} \left[ (\Delta_{t}^{\pi_{t}, b_{1:t}, \cdot, \cdot, \cdot})^{2} | \mathcal{F}_{t-1} \right] \leq H^{2} \prod_{k \in W} \pi_{i,k}^{t}(b_{k}|x_{k}) \prod_{k \in [h']} \pi_{i,k}^{t,h}(a_{k}|x_{k}) \prod_{k \in W(h)} \pi_{i,k}^{t,h}(a_{k}|x_{k}).
\]

Here, (i) is because \( \prod_{k \in [h']} \pi_{i,k}^{t,h}(a_{k}|x_{k}) = \prod_{k \in [h']} \pi_{i,k}^{t}(b_{k}|x_{k}) \).

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

\[
\sum_{t=1}^{T} \Delta_{t}^{\pi_{t}, b_{1:t}, \cdot, \cdot, \cdot} \leq \lambda H^{2} \prod_{k \in W(h)} \pi_{i,k}^{t,h}(a_{k}|x_{k}) \sum_{t=1}^{T} \prod_{k \in [h']} \pi_{i,k}^{t}(b_{k}|x_{k}) p_{1:h}^{t}(x_{i,h}) + C \log(X_{i}A_{i}/p) \lambda,
\]

where \( C > 0 \) is some absolute constant. Plugging this bound into \( \text{BIAS}_{T, h}^{\cdot, \cdot, \cdot} \) yields that, for all \( h \in [H] \) and \( \phi \in \Phi_{K}^{h} \),

\[
\text{BIAS}_{T, h}^{\cdot, \cdot, \cdot} \leq \max_{\phi \in \Phi_{K}^{h}} \sum_{x_{i,h}, b_{1:h-1}} \sum_{b_{1:h-1}} \delta^{K}(a_{1:h-1}, b_{1:h-1}) \prod_{k \in W(h)} \pi_{i,k}^{h}(a_{k}|x_{k}) \prod_{k \in [h']} \pi_{i,k}^{t}(b_{k}|x_{k}) p_{1:h}^{t}(x_{i,h}) \left( \lambda H^{2} \prod_{k \in W(h)} \pi_{i,k}^{t,h}(a_{k}|x_{k}) \sum_{t=1}^{T} \prod_{k \in [h']} \pi_{i,k}^{t}(b_{k}|x_{k}) p_{1:h}^{t}(x_{i,h}) + C \log(X_{i}A_{i}/p) \lambda \right) \]

\[
\leq \max_{\phi \in \Phi_{K}^{h}} \lambda H^{2} \sum_{x_{i,h}, b_{1:h-1}} \sum_{b_{1:h-1}} \delta^{K}(a_{1:h-1}, b_{1:h-1}) \prod_{k \in W(h)} \phi_{k}(a_{k}|x_{k}, b_{1:k,h'}) \sum_{t=1}^{T} \prod_{k \in [h']} \pi_{i,k}^{t}(b_{k}|x_{k}) p_{1:h}^{t}(x_{i,h}) + \max_{\phi \in \Phi_{K}^{h}} \frac{C \log(X_{i}A_{i}/p)}{\lambda} \sum_{x_{i,h}, b_{1:h-1}} \sum_{b_{1:h-1}} \delta^{K}(a_{1:h-1}, b_{1:h-1}) \prod_{k \in W(h)} \phi_{k}(a_{k}|x_{k}, b_{1:k,h'}) \prod_{k \in [h']} \pi_{i,k}^{t}(b_{k}|x_{k}) p_{1:h}^{t}(x_{i,h}) \leq 1.
\]
So we have

\[ \sum_{a_h, a_{1:h-1}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \sum_{t=1}^{T} \prod_{k \in [h]} \pi^t_{i,k}(b_k|x_k)p^{x_t}_{i,h}(x_{i,h}) \leq T. \]

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

\[ \sum_{a_h, a_{1:h-1}} \sum_{b_{1:h-1}} \delta^K(a_{1:h-1}, b_{1:h-1}) \prod_{k=1}^{h} \phi_k(a_k|x_k, b_{1:k\wedge h'}) \prod_{k \in [h]} \pi^t_{i,k}(b_k|x_k) \leq X_{i,h} \left( \frac{h - 1}{K} \right) A_i^{K_A H}. \]

Plugging these bounds into BIAS\(_{2,h}^{T,ext}\) yields that, with probability at least 1 – 8/p, for all \( h \in [H] \),

\[ \text{BIAS}_{2,h}^{T,ext} \leq \lambda H^2 T + \frac{C \log(X_i A_i/p)}{\lambda} X_{i,h} \left( \frac{h - 1}{K} \right) A_i^{K_A H}. \]

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

\[ \sum_{h=1}^{H} \text{BIAS}_{2,h}^{T,ext} \leq \lambda H^3 T + \frac{C \log(X_i A_i/p)}{\lambda} \sum_{h=1}^{H} X_{i,h} \left( \frac{h - 1}{K} \right) A_i^{K_A H}, \]

for all \( \lambda \in (0, 1/H] \). Choose \( \lambda = \min \left\{ \frac{1}{H}, \sqrt{\frac{C X_i (K_A H)}{H} A_i^{K_A H} \log(X_i A_i/p)} \right\} \), we obtain the bound

\[ \sum_{t=1}^{T} \text{BIAS}_{2,h}^{T,ext} \leq O \left( \sqrt{H^3 \left( \frac{H}{K} \right) A_i^{K_A H} X_i T_{1:t} + H \left( \frac{H}{K} \right) A_i^{K_A H} X_i T_{1:t}} \right), \]

where \( t = \log(8X_i A_i/p) \) is a log factor. \( \square \)

### 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 \( \{\pi^t_{i,h}\}_{h \in [H]} \) depend on the number of children \( C_h(x_{i,h}, a_{i,h}) \) 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 $\varepsilon$-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).