NEAR-OPTIMAL LEARNING OF EXTENSIVE-FORM GAMES WITH IMPERFECT INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper resolves the open question of designing near-optimal algorithms for learning imperfect-information extensive-form games from bandit feedback. We present the first line of algorithms that require only $\widetilde{\mathcal{O}}((XA+YB)/\varepsilon^2)$ episodes of play to find an ε-approximate Nash equilibrium in two-player zero-sum games, where X, Y are the number of information sets and A, B are the number of actions for the two players. This improves upon the best known sample complexity of $\widetilde{\mathcal{O}}((X^2A+Y^2B)/\varepsilon^2)$ by a factor of $\widetilde{\mathcal{O}}(\max\{X,Y\})$, and matches the information-theoretic lower bound up to logarithmic factors. We achieve this sample complexity by two new algorithms: Balanced Online Mirror Descent, and Balanced Counterfactual Regret Minimization. Both algorithms rely on novel approaches of integrating *balanced exploration policies* into their classical counterparts. We also extend our results to learning Coarse Correlated Equilibria in multi-player general-sum games.

1 INTRODUCTION

Imperfect Information Games—games where players can only make decisions based on partial information about the true underlying state of the game—constitute an important challenge for modern artificial intelligence. The celebrated notion of Imperfect-Information Extensive-Form games (IIEFGs) (Kuhn, 1953) offers a formulation for games with both imperfect information and sequential play. IIEFGs have been widely used for modeling real-world imperfect information games such as Poker (Heinrich et al., 2015; Moravčík et al., 2017; Brown & Sandholm, 2018), Bridge (Tian et al., 2020), Scotland Yard (Schmid et al., 2021), etc, and achieving strong performances therein.

A central question in IIEFGs is the problem of finding a Nash equilibrium (NE) (Nash, 1950) in a two-player zero-sum IIEFG with perfect recall. There is an extensive line of work for solving this problem with full knowledge of the game (or full feedback), by either reformulating as a linear program (Koller & Megiddo, 1992; Von Stengel, 1996; Koller et al., 1996), first-order optimization methods (Hoda et al., 2010; Kroer et al., 2015; 2018; Munos et al., 2020; Lee et al., 2021), or Counterfactual Regret Minimization (Zinkevich et al., 2007; Lanctot et al., 2009; Johanson et al., 2012; Tammelin, 2014; Schmid et al., 2019; Burch et al., 2019).

However, in the more challenging bandit feedback setting where the game is not known and can only be learned from random observations by repeated playing, the optimal sample complexity (i.e., the number of episodes required to play) for learning an NE in IIEFGs remains open. Various approaches have been proposed recently for solving this, including model-based exploration (Zhou et al., 2019; Zhang & Sandholm, 2021), Online Mirror Descent with loss estimation (Farina et al., 2021; Kozuno et al., 2021), and Monte-Carlo Counterfactual Regret Minimization (MCCFR) (Lanctot et al., 2009; Farina et al., 2020c; Farina & Sandholm, 2021). In a two-player zero-sum IIEFG with X, Y information sets (infosets) and A, B actions for the two players respectively, the current best sample complexity for learning an ε-approximate NE is $\widetilde{\mathcal{O}}((X^2A+Y^2B)/\varepsilon^2)$ achieved by a sample-based variant of Online Mirror Descent with implicit exploration (Kozuno et al., 2021). However, this sample complexity scales quadratically in X, Y and still has a gap from the information-theoretic lower bound $\Omega((XA+YB)/\varepsilon^2)$ which only scales linearly. This gap is especially concerning from a practical point of view as the number of infosets is often the dominating measure of the game size in large real-world IIEFGs (Johanson, 2013).

Algorithm	OMD	CFR	Sample Complexity
Zhang & Sandholm (2021)	- (model-based)		$\widetilde{\mathcal{O}}\left(S^2AB/\varepsilon^2\right)$
Farina & Sandholm (2021)		✓	$\widetilde{\mathcal{O}}(\text{poly}\left(X,Y,A,B\right)/\varepsilon^4)$
Farina et al. (2021)	✓		$\widetilde{\mathcal{O}}\left(\left(X^4A^3+Y^4B^3\right)/\varepsilon^2\right)$
Kozuno et al. (2021)	✓		$\widetilde{\mathcal{O}}\left(\left(X^2A+Y^2B\right)/\varepsilon^2\right)$
Balanced OMD (Algorithm 1)	✓		$\widetilde{\mathcal{O}}\left(\left(XA+YB\right)/\varepsilon^2\right)$
Balanced CFR (Algorithm 2)		✓	$\widetilde{\mathcal{O}}\left(\left(XA+YB\right)/\varepsilon^2\right)$
Lower bound (Theorem F.1)	-	-	$\Omega\left(\left(XA+YB\right)/\varepsilon^2\right)$

Table 1: Sample complexity (number of episodes required) for learning ε-NE in IIEFGs from bandit feedback.

In this paper, we resolve this open question by presenting the first line of algorithms that achieve $\widetilde{\mathcal{O}}((XA+YB)/\varepsilon^2)$ sample complexity for learning ε-NE in an IIEFG. Our contributions can be summarized as follows.

• We design a new algorithm Balanced Online Mirror Descent (Balanced OMD) that achieves $\widetilde{\mathcal{O}}(\sqrt{XAT})$ regret for the max player against adversarial opponents, and learns an ε-NE within $\widetilde{\mathcal{O}}((XA+YB)/\varepsilon^2)$ episodes of play when run by both players in a self-play fashion (Section 3). These improve over the best existing results by a factor of $\sqrt{X}$ and $\max\{X,Y\}$ respectively, and match the information-theoretic lower bounds (Appendix F) up to $\text{poly}(H)$ and logarithmic factors. The main feature within Balanced OMD is a new *balanced dilated KL* as the distance function in its mirror descent step.

• We design another new algorithm Balanced Counterfactual Regret Minimization (Balanced CFR) that also achieves $\widetilde{\mathcal{O}}((XA+YB)/\varepsilon^2)$ sample complexity for learning an ε-NE (Section 4). Balanced CFR can be seen as an instantiation of the MCCFR framework that integrates the *balanced exploration policies* within both the sampling and the local regret minimization steps.

• We extend our results to multi-player general-sum IIEFGs, where we show that both Balanced OMD and Balanced CFR can learn an approximate Normal-Form Coarse Correlated Equilibrium (NFCCE) sample-efficiently when run by all players simultaneously via self-play (Appendix I).

Review of additional related work can be found in Appendix A.

2 Preliminaries

We consider two-player zero-sum IIEFGs using the formulation via Partially Observable Markov Games (POMGs), following (Kozuno et al., 2021). In the following, $\Delta(\mathcal{A})$ denotes the probability simplex over a set $\mathcal{A}$.

Partially observable Markov games We consider finite-horizon, tabular, two-player zero-sum Markov Games with partial observability, which can be described as a tuple $\text{POMG}(H,\mathcal{S},\mathcal{X},\mathcal{Y},\mathcal{A},\mathcal{B},\mathbb{P},r)$, where H is the horizon length; $\mathcal{S}=\bigcup_{h\in[H]}\mathcal{S}_h$ is the (underlying) state space with $|\mathcal{S}_h|=S_h$ and $\sum_{h=1}^H S_h=S$; $\mathcal{X}=\bigcup_{h\in[H]}\mathcal{X}_h$ is the space of information sets (henceforth *infosets*) for the *max-player* with $|\mathcal{X}_h|=X_h$ and $X:=\sum_{h=1}^H X_h$. At any state $s_h\in\mathcal{S}_h$, the max-player only observes the infoset $x_h=x(s_h)\in\mathcal{X}_h$, where $x:\mathcal{S}\to\mathcal{X}$ is the emission function for the max-player; $\mathcal{Y}=\bigcup_{h\in[H]}\mathcal{Y}_h$ is the space of infosets for the *min-player* with $|\mathcal{Y}_h|=Y_h$ and $Y:=\sum_{h=1}^H Y_h$. An infoset y_h and the emission function $y:\mathcal{S}\to\mathcal{Y}$ are defined similarly. $\mathcal{A}$, $\mathcal{B}$ are the action spaces for the max-player and min-player respectively, with $|\mathcal{A}|=A$ and $|\mathcal{B}|=B$[1]. $\mathbb{P}=\{p_0(\cdot)\in\Delta(\mathcal{S}_1)\}\cup\{p_h(\cdot|s_h,a_h,b_h)\in\Delta(\mathcal{S}_{h+1})\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B},\ h\in[H-1]}$ are the transition probabilities, where $p_1(s_1)$ is the probability of the initial state being s_1, and $p_h(s_{h+1}|s_h,a_h,b_h)$ is the probability of transitting to s_{h+1} given state-action (s_h,a_h,b_h) at step

[1]While this assumes the action space at each infoset have equal sizes, our results can be extended directly to the case where each infoset has its own action space with (potentially) unequal sizes.

h; $r = \{r_h(s_h, a_h, b_h) \in [0,1]\}_{(s_h,a_h,b_h)\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{B}}$ are the (random) reward functions with mean $\overline{r}_h(s_h, a_h, b_h)$.

Policies, value functions As we consider partially observability, each player's policy can only depend on the infoset rather than the underlying state. A policy for the max-player is denoted by $\mu = \{\mu_h(\cdot|x_h) \in \Delta(\mathcal{A})\}_{h\in[H],x_h\in\mathcal{X}_h}$, where $\mu_h(a_h|x_h)$ is the probability of taking action $a_h \in \mathcal{A}$ at infoset $x_h \in \mathcal{X}_h$. Similarly, a policy for the min-player is denoted by $\nu = \{\nu_h(\cdot|y_h) \in \Delta(\mathcal{B})\}_{h\in[H],y_h\in\mathcal{Y}_h}$. A trajectory for the max player takes the form $(x_1, a_1, r_1, x_2, \ldots, x_H, a_H, r_H)$, where $a_h \sim \mu_h(\cdot|x_h)$, and the rewards and infoset transitions depend on the (unseen) opponent's actions and underlying state transition. The overall game value for any (product) policy (μ, ν) is denoted by $V^{\mu,\nu} := \mathbb{E}_{\mu,\nu}\left[\sum_{h=1}^{H} r_h(s_h, a_h, b_h)\right]$. The max-player aims to maximize the value, whereas the min-player aims to minimize the value.

Tree structure and perfect recall We use a POMG with tree structure and the perfect recall assumption as our formulation for IIEFGs, following (Kozuno et al., 2021). We assume that our POMG has a *tree structure*: For any h and $s_h \in \mathcal{S}_h$, there exists a unique history $(s_1, a_1, b_1, \ldots, s_{h-1}, a_{h-1}, b_{h-1})$ of past states and actions that leads to s_h. We also assume that both players have *perfect recall*: For any h and any infoset $x_h \in \mathcal{X}_h$ for the max-player, there exists a unique history $(x_1, a_1, \ldots, x_{h-1}, a_{h-1})$ of past infosets and max-player actions that leads to x_h (and similarly for the min-player). We further define $\mathcal{C}_{h'}(x_h, a_h) \subset \mathcal{X}_{h'}$ to be the set of all infosets in the h'-the step that are reachable from (x_h, a_h), and define $\mathcal{C}_{h'}(x_h) = \cup_{a_h\in\mathcal{A}}\mathcal{C}_{h'}(x_h, a_h)$. Finally, define $\mathcal{C}(x_h, a_h) := \mathcal{C}_{h+1}(x_h, a_h)$ as a shorthand for immediately reachable infosets.

With the tree structure and perfect recall, under any product policy (μ, ν), the probability of reaching state-action (s_h, a_h, b_h) at step h takes the form

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

where we have defined the sequence-form transition probability as

$$p_{1:h}(s_h) := p_0(s_1) \prod_{h'\le h-1} p_{h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'}),$$

where $\{s_{h'}, a_{h'}, b_{h'}\}_{h'\le h-1}$ are the histories uniquely determined from s_h by the tree structure, and the *sequence-form policies* as

$$\mu_{1:h}(x_h, a_h) := \prod_{h'=1}^{h} \mu_{h'}(a_{h'}|x_{h'}), \quad \nu_{1:h}(y_h, b_h) := \prod_{h'=1}^{h} \nu_{h'}(b_{h'}|y_{h'}),$$

where $x_{h'} = x(s_{h'})$ and $y_{h'} = y(s_{h'})$ are the infosets for the two players (with $\{x_{h'}, a_{h'}\}_{h\le h-1}$ are uniquely determined by x_h by perfect recall, and similar for $\{y_{h'}, b_{h'}\}_{h\le h-1}$).

We let $\Pi_{\max}$ denote the set of all possible policies for the max player ($\Pi_{\min}$ for the min player). In the sequence form representation, $\Pi_{\max}$ is a convex compact subset of $\mathbb{R}^{XA}$ specified by the constraints $\mu_{1:h}(x_h, a_h) \ge 0$ and $\sum_{a_h\in\mathcal{A}} \mu_{1:h}(x_h, a_h) = \mu_{1:h-1}(x_{h-1}, a_{h-1})$ for all (h, x_h, a_h), where (x_{h-1}, a_{h-1}) is the unique pair of prior infoset and action that reaches x_h (understanding $\mu_0(x_0, a_0) = \mu_0(\emptyset) = 1$).

Regret and Nash Equilibrium We consider two standard learning goals: Regret and Nash Equilibrium. For the regret, we focus on the max-player, and assume there is an arbitrary (potentially adversarial) opponent as the min-player who may determine her policy ν^t based on all past information (including knowledge of μ^t) before the t-th episode starts. Then, the two players play the t-th episode jointly using (μ^t, ν^t). The goal for the max-player's is to design policies $\{\mu^t\}_{t=1}^{T}$ that minimizes the regret against the best fixed policy in hindsight:

$$\mathfrak{R}^T := \max_{\mu^\dagger\in\Pi_{\max}} \sum_{t=1}^{T} \left(V^{\mu^\dagger,\nu^t} - V^{\mu^t,\nu^t}\right). \tag{2}$$

We say a product policy (μ, ν) is an ε-approximate Nash equilibrium (ε-NE) if

$$\mathrm{NEGap}(\mu, \nu) := \max_{\mu^\dagger\in\Pi_{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger\in\Pi_{\min}} V^{\mu,\nu^\dagger} \le \varepsilon,$$

i.e. μ and ν are each other's ε-approximate best response. Using online-to-batch conversion, it is a standard result that sublinear regret for both players ensures that the pair of average policies $(\overline{\mu}, \overline{\nu})$ is an approximate NE (see e.g. (Kozuno et al., 2021, Theorem 1)):

Proposition 1 (Regret to Nash conversion)**.** *For any sequence of policies* $\{\mu^t\}_{t=1}^T \in \Pi_{\max}$ *and* $\{\nu^t\}_{t=1}^T \in \Pi_{\min}$*, the average policies* $\overline{\mu} := \frac{1}{T}\sum_{t=1}^T \mu^t$ *and* $\overline{\nu} := \frac{1}{T}\sum_{t=1}^T \nu^t$ *(averaged in the sequence form, cf.* (17)*) satisfy*

$$\mathrm{NEGap}(\overline{\mu}, \overline{\nu}) = (\mathfrak{R}_{\max}^T + \mathfrak{R}_{\min}^T)/T,$$

where $\mathfrak{R}_{\max}^T := \max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^T (V^{\mu^\dagger, \nu^t} - V^{\mu^t, \nu^t})$ *and* $\mathfrak{R}_{\min}^T := \max_{\nu^\dagger \in \Pi_{\min}} \sum_{t=1}^T (V^{\mu^t, \nu^t} - V^{\mu^t, \nu^\dagger})$ *denote the regret for the two players respectively.*

Therefore, an approximate NE can be learned by letting both players play some sublinear regret algorithm against each other in a self-play fashion.

Bandit feedback Throughout this paper, we consider the interactive learning (exploration) setting with bandit feedback, where the max-player determines the policy μ^t, the opponent determines ν^t (either adversarially or by running some learning algorithm, depending on the context) unknown to the max-player, and the two players play an episode of the game using policy (μ^t, ν^t). The max player observes the trajectory $(x_1^t, a_1^t, r_1^t, \ldots, x_H^t, a_H^t, r_H^t)$ of her own infosets and rewards, but not the opponent's infosets, actions, or the underlying states.

2.1 Conversion to online linear regret minimization

The reaching probability decomposition (1) implies that the value function $V^{\mu,\nu}$ is *bilinear* in (the sequence form of) (μ, ν). Thus, fixing a sequence of opponent's policies $\{\nu^t\}_{t=1}^T$, we have the linear representation

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

Therefore, defining the *loss function* for round t as

$$\ell_h^t(x_h, a_h) := \sum_{s_h\in x_h, b_h\in\mathcal{B}} p_{1:h}(s_h)\nu_{1:h}^t(y(s_h), b_h)(1 - \overline{r}_h(s_h, a_h, b_h)) \tag{3}$$

the regret $\mathfrak{R}^T$ (2) can be written as

$$\mathfrak{R}^T = \max_{\mu^\dagger\in\Pi_{\max}} \sum_{t=1}^T \left\langle \mu^t - \mu^\dagger, \ell^t \right\rangle, \tag{4}$$

where the inner product $\langle \mu, \ell^t \rangle := \sum_{h=1}^H \sum_{x_h,a_h} \mu_{1:h}(x_h, a_h)\ell_h^t(x_h, a_h)$ for any $\mu \in \Pi_{\max}$.

2.2 Balanced exploration policy

Our algorithms make crucial use of the following balanced exploration policies.

Definition 2 (Balanced exploration policy)**.** *For any* $1 \le h \le H$*, the (max-player's)* balanced exploration policy for layer h*, denoted as* $\mu^{\star,h} \in \Pi_{\max}$*, is defined as*

$$\mu_{h'}^{\star,h}(a_{h'}|x_{h'}) := \begin{cases} |\mathcal{C}_h(x_{h'}, a_{h'})|/|\mathcal{C}_h(x_{h'})|, & h' \in \{1, \ldots, h-1\}, \\ 1/A, & h' \in \{h, \ldots, H\}. \end{cases} \tag{5}$$

In words, at time steps $h' \le h-1$*, the policy* $\mu^{\star,h}$ *plays actions proportionally to their number of descendants* within the h-th layer *of the game tree. At time steps* $h' \ge h$*, it plays the uniform policy.*

Note that there are H such balanced policies, one for each layer $h \in [H]$. The balanced policy for layer $h = H$ is equivalent to the balanced strategy of Farina et al. (2020c) (cf. their Section 4.2 and Appendix A.3) which plays actions proportionally to their number descendants within the last (terminal) layer. The balanced policies for layers $h \le H-1$ generalize theirs by also counting the number of descendants within earlier layers. We remark in passing that the key feature of $\mu^{\star,h}$ for our analyses is its *balancing property*, which we state in Lemma D.4.

Requirement on knowing the tree structure The construction of $\mu^{\star,h}$ requires knowing the number of descendants $|\mathcal{C}_h(x_{h'}, a_{h'})|$, which depends on the *structure*[2] *of the game tree* for the max player. Therefore, our algorithms that use $\mu^{\star,h}$ requires knowing this tree structure beforehand. Although there exist algorithms that do not require knowing such tree structure beforehand (Zhang & Sandholm, 2021; Kozuno et al., 2021), this requirement is relatively mild as the structure can be extracted efficiently from just one tree traversal. We also remark our algorithms using the balanced policies do not impose any additional requirements on the game tree, such as the existence of a policy with lower bounded reaching probabilities at all infosets.

3 Online Mirror Descent

We now present our first algorithm Balanced Online Mirror Descent (Balanced OMD) and its theoretical guarantees.

3.1 Balanced dilated KL

At a high level, OMD algorithms work by designing loss estimators (e.g. using importance weighting) and solving a regularized optimization problem in each round involving the loss estimator and a distance function as the regularizer. OMD has been successfully deployed for solving IIEFGs by using various *dilated distance generating functions* over the policy set $\Pi_{\max}$ (Kroer et al., 2015).

The main ingredient of our algorithm is the *balanced dilated KL*, a new distance measure between policies in IIEFGs.

Definition 3 (Balanced dilated KL)**.** *The balanced dilated KL distance between two policies $\mu, \nu \in \Pi_{\max}$ is defined as*

$$\mathrm{D}^{\mathrm{bal}}(\mu \| \nu) := \sum_{h=1}^{H} \sum_{x_h, a_h} \frac{\mu_{1:h}(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \log \frac{\mu_h(a_h|x_h)}{\nu_h(a_h|x_h)}. \tag{6}$$

The balanced dilated KL is a reweighted version of the *dilated KL* (a.k.a. the *dilated entropy distance-generating function*) that has been widely used for solving IIEFGs (Hoda et al., 2010; Kroer et al., 2015):

$$\mathrm{D}(\mu \| \nu) = \sum_{h=1}^{H} \sum_{x_h, a_h} \mu_{1:h}(x_h, a_h) \log \frac{\mu_h(a_h|x_h)}{\nu_h(a_h|x_h)}. \tag{7}$$

Compared with (7), our balanced dilated KL (6) introduces an additional reweighting term $1/\mu^{\star,h}_{1:h}(x_h, a_h)$ that depends on the balanced exploration policy $\mu^{\star,h}$ (5). This reweighting term is in general different for each (x_h, a_h), which will introduce a balancing effect into our algorithm.

3.2 Algorithm and theoretical guarantee

We now describe our Balanced OMD algorithm in Algorithm 1. Our algorithm is a variant of the IXOMD algorithm of Kozuno et al. (2021) by using the balanced dilated KL. At a high level, it consists of the following steps:

• Line 3 & 5 (Sampling): Play an episode using policy μ^t (against the opponent ν^t) and observe the trajectory. Then construct the loss estimator using importance weighting and IX bonus (Neu, 2015):

$$\widetilde{\ell}^t_h(x_h, a_h) := \frac{\mathbf{1}\left\{(x^t_h, a^t_h) = (x_h, a_h)\right\} \cdot (1 - r^t_h)}{\mu^t_{1:h}(x_h, a_h) + \gamma \mu^{\star,h}_{1:h}(x_h, a_h)}. \tag{8}$$

Note that the IX bonus $\gamma \mu^{\star,h}_{1:h}(x_h, a_h)$ on the denominator makes (8) a slightly downward biased estimator of the true loss $\ell^t_h(x_h, a_h)$ defined in (3).

• Line 6 (Update policy): Update μ^{t+1} by OMD with loss estimator $\widetilde{\ell}^t$ and the balanced dilated KL distance function. Due to the sparsity of $\widetilde{\ell}^t$, this update admits an efficient implementation that updates the conditional form $\mu^t_h(\cdot|x_h)$ at the visited infoset $x_h = x^t_h$ only (described in Algorithm 5).

We are now ready to present the theoretical guarantees for the Balanced OMD algorithm.

[2] By this "structure" we refer to the parenting structure of the game tree only (which x_{h+1} is reachable from which (x_h, a_h)), not the transition probabilities and rewards.

Algorithm 1 Balanced OMD (max-player)

Require: Learning rate $\eta > 0$; IX parameter $\gamma > 0$.

1: Initialize $\mu_h^1(a_h|x_h) \leftarrow 1/A_h$ for all (h, x_h, a_h).

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

3: Play an episode using μ^t, observe a trajectory

$$(x_1^t, a_1^t, r_1^t, \ldots, x_H^t, a_H^t, r_H^t).$$

4: **for** $h = H, \ldots, 1$ **do**

5: Construct loss estimator $\left\{\widetilde{\ell}_h^t(x_h, a_h)\right\}_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}$ by

$$\widetilde{\ell}_h^t(x_h, a_h) \leftarrow \frac{\mathbf{1}\left\{(x_h^t, a_h^t) = (x_h, a_h)\right\} \cdot (1 - r_h^t)}{\mu_{1:h}^t(x_h, a_h) + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)}.$$

6: Update policy

$$\mu^{t+1} \leftarrow \operatorname*{arg\,min}_{\mu\in\Pi_{\max}} \left\langle \mu, \widetilde{\ell}^t \right\rangle + \frac{1}{\eta}\mathrm{D}^{\mathrm{bal}}(\mu\|\mu^t) \tag{9}$$

using the efficient implementation in Algorithm 5.

Theorem 4 (Regret bound for Balanced OMD)**.** *Algorithm 1 with learning rate* $\eta = \sqrt{XA\log A/(H^3T)}$ *and IX parameter* $\gamma = \sqrt{XA\iota/(HT)}$ *achieves the following regret bound with probability at least* $1-\delta$*:*

$$\mathfrak{R}^T \le \mathcal{O}\Big(\sqrt{H^3XAT\iota}\Big),$$

where $\iota := \log(3HXA/\delta)$ *is a log factor.*

Letting both players run Algorithm 1, the following corollary for learning NE follows immediately from the regret-to-Nash conversion (Proposition 1).

Corollary 5 (Learning NE using Balanced OMD)**.** *Suppose both players run Algorithm 1 (and its min player's version) against each other for* T *rounds, with choices of* η, γ *specified in Theorem 4. Then, for any* $\varepsilon > 0$*, the average policy* $(\overline{\mu}, \overline{\nu}) = (\frac{1}{T}\sum_{t=1}^T \mu^t, \frac{1}{T}\sum_{t=1}^T \nu^t)$ *achieves* $\mathrm{NEGap}(\overline{\mu}, \overline{\nu}) \le \varepsilon$ *with probability at least* $1-\delta$*, as long as the number of episodes*

$$T \ge \mathcal{O}\big(H^3(XA+YB)\iota/\varepsilon^2\big),$$

where $\iota := \log(3H(XA+YB)/\delta)$ *is a log factor.*

Theorem 4 and Corollary 5 are the first to achieve $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\cdot\sqrt{XAT})$ regret and $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\cdot(XA+YB)/\varepsilon^2)$ sample complexity for learning an ε-approximate NE for IIEFGs. Notably, the sample complexity scales only linearly in X, Y and improves significantly over the best known $\widetilde{\mathcal{O}}((X^2A+Y^2B)/\varepsilon^2))$ achieved by the IXOMD algorithm of (Kozuno et al., 2021) by a factor of $\max\{X, Y\}$. The proof of Theorem 4 (and a brief overview) can be found in Appendix E.2.

Lower bound We accompany our results with information-theoretic lower bounds (stated in Theorem F.1) showing that our $\widetilde{\mathcal{O}}(\sqrt{H^3XAT})$ regret and $\widetilde{\mathcal{O}}(H^3(XA+YB)/\varepsilon^2)$ sample complexity are both near-optimal modulo $\mathrm{poly}(H)$ and log factors.

4 Counterfactual Regret Minimization

Counterfactual Regret Minimization (CFR) (Zinkevich et al., 2007) is another widely used class of algorithms for solving IIEFGs. In this section, we present a new variant Balanced CFR that also achieves sharp sample complexity guarantees.

Different from OMD, CFR-type algorithms maintain a "local" regret minimizer at each infoset x_h that aims to minimize the *immediate counterfactual regret* at that infoset:

$$\mathfrak{R}_h^{\mathrm{imm},T}(x_h) := \max_{\mu\in\Delta(\mathcal{A})} \sum_{t=1}^T \left\langle \mu_h^t(\cdot|x_h) - \mu(\cdot), L_h^t(x_h, \cdot)\right\rangle,$$

where $L_h^t(x_h, a_h)$ is the *counterfactual loss function*

$$L_h^t(x_h, a_h) := \ell_h^t(x_h, a_h) + \sum_{h'=h+1}^{H} \sum_{(x_{h'}, a_{h'}) \in \mathcal{C}_{h'}(x_h, a_h) \times \mathcal{A}} \mu_{(h+1):h'}^t(x_{h'}, a_{h'}) \ell_{h'}^t(x_{h'}, a_{h'}). \quad (10)$$

Controlling all the immediate counterfactual regrets $\mathfrak{R}_h^{\mathrm{imm},T}(x_h)$ will also control the overall regret of the game $\mathfrak{R}^T$, as guaranteed by the *counterfactual regret decomposition* (Zinkevich et al., 2007) (see also our Lemma G.1).

4.1 Algorithm description

Our Balanced CFR algorithm, described in Algorithm 2, can be seen as an instantiation of the Monte-Carlo CFR (MCCFR) framework (Lanctot et al., 2009) that incorporates the balanced policies in its sampling procedure. Algorithm 2 requires regret minimization algorithms R_{x_h} for each x_h as its input, and performs the following steps in each round:

- Line 4-6 (Sampling): Play H episodes using policies $\{\mu^{t,(h)}\}_{h \in [H]}$, where each $\mu^{t,(h)} = (\mu_{1:h}^{\star,h} \mu_{h+1:H}^t)$ is a *mixture* of the balanced exploration policy $\mu^{\star,h}$ with the current maintained policy μ^t over time steps. Then, compute $\widetilde{L}_h^t(x_h, a_h)$ by (12) that are importance-weighted unbiased estimators of the true counterfactual loss $L_h^t(x_h, a_h)$ in (10).

- Line 8 (Update regret minimizers): For each (h, x_h), send the loss estimators $\{\widetilde{L}_h^t(x_h, a)\}_{a \in \mathcal{A}}$ to the local regret minimizer R_{x_h}, and obtain the updated policy $\mu_h^{t+1}(\cdot|x_h)$.

Similar as existing CFR-type algorithms, Balanced CFR has the flexibility of allowing different choices of regret minimization algorithms as R_{x_h}. We will consider two concrete instantiations of R_{x_h} as Hedge and Regret Matching in the following subsection.

4.2 Theoretical guarantee

To obtain a sharp guarantee for Balanced CFR, we first instantiate R_{x_h} as the Hedge algorithm (a.k.a. Exponential Weights, or mirror descent with the entropic regularizer; cf. Algorithm 3). Specifically, we let each R_{x_h} be the Hedge algorithm with learning rate $\eta \mu_{1:h}^{\star,h}(x_h, a)$[3]. With this choice, Line 8 of Algorithm 2 takes the following explicit form:

$$\mu_h^{t+1}(a|x_h) \propto_a \mu_h^t(a|x_h) \cdot e^{-\eta \mu_{1:h}^{\star,h}(x_h,a) \cdot \widetilde{L}_h^t(x_h,a)}. \quad (11)$$

We are now ready to present the theoretical guarantees for the Balanced CFR algorithm.

Theorem 6 ("Regret" bound for Balanced CFR)**.** *Suppose the max player plays Algorithm 2 where each R_{x_h} is instantiated as the Hedge algorithm (11) with $\eta = \sqrt{XA\iota/(H^3T)}$. Then, the policies μ^t achieve the following "regret" bound with probability at least $1-\delta$:*

$$\widetilde{\mathfrak{R}}^T := \max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \langle \mu^t - \mu^\dagger, \ell^t \rangle \le \mathcal{O}(\sqrt{H^3 XAT\iota}),$$

where $\iota = \log(10XA/\delta)$ is a log factor.

The $\widetilde{\mathcal{O}}(\sqrt{H^3XAT})$ "regret" achieved by Balanced CFR matches that of Balanced OMD. However, we emphasize that the quantity $\widetilde{\mathfrak{R}}^T$ is *not strictly speaking a regret*, as it measures performance of the policy $\{\mu^t\}$ *maintained* in the Balanced CFR algorithm, not the sampling policy $\mu^{t,(h)}$ that the Balanced CFR algorithm have *actually played*. Nevertheless, we remark that such a form of "regret" bound is the common type of guarantee for all existing MCCFR type algorithms (Lanctot et al., 2009; Farina et al., 2020c).

Self-play of Balanced CFR Balanced CFR can also be turned into a PAC algorithm for learning ε-NE, by letting the two players play Algorithm 2 against each other for T rounds of self-play using the following protocol: Within each round, the max player plays policies $\{\mu^{t,(h)}\}_{h=1}^H$ while the min player plays the fixed policy ν^t; then symmetrically the min player plays $\{\nu^{t,(h)}\}_{h=1}^H$ while the

[3]Note that this quantity depends on x_h but not a.

Algorithm 2 Balanced CFR (max-player)

Require: Regret minimization algorithm R_{x_h} for all (h, x_h).

Initialize policy $\mu_h^1(a_h|x_h) \leftarrow 1/A$ for all (h, x_h, a_h).
for round $t = 1, \dots, T$ **do**
  for $h = 1, \dots, H$ **do**
    Set policy $\mu^{t,(h)} \leftarrow (\mu_{1:h}^{\star,h} \mu_{h+1:H}^t)$.
    Play an episode using $\mu^{t,(h)} \times \nu^t$, observe a trajectory

$$(x_1^{t,(h)}, a_1^{t,(h)}, r_1^{t,(h)}, \cdots, x_H^{t,(h)}, a_H^{t,(h)}, r_H^{t,(h)}).$$

    Compute loss estimators for all (h, x_h, a_h):

$$\widetilde{L}_h^t(x_h, a_h) := \frac{\mathbf{1}\left\{(x_h^{t,(h)}, a_h^{t,(h)}) = (x_h, a_h)\right\}}{\mu_{1:h}^{\star,h}(x_h, a_h)} \left(H - h + 1 - \sum_{h'=h}^{H} r_{h'}^{t,(h)} \right). \tag{12}$$

  for all $h \in [H]$ and $x_h \in \mathcal{X}_h$ **do**
    Update the regret minimizer at x_h and obtain policy:

$$\mu_h^{t+1}(\cdot|x_h) \leftarrow R_{x_h}.\textsc{Update}(\{\widetilde{L}_h^t(x_h, a)\}_{a \in \mathcal{A}}). \tag{13}$$

max player plays the fixed μ^t. Overall, each round plays $2H$ episodes. Theorem 6 directly implies the following corollary for the above self-play algorithm on learning ε-NE, by the regret-to-Nash conversion (Proposition 1).

Corollary 7 (Learning NE using Balanced CFR)**.** *Let both players play Algorithm 2 in a self-play fashion against each other for T rounds, where each R_{x_h} is instantiated as the Hedge algorithm (11) with η specified in Theorem 6. Then, for any $\varepsilon > 0$, the average policy $(\overline{\mu}, \overline{\nu}) = (\frac{1}{T}\sum_{t=1}^T \mu^t, \frac{1}{T}\sum_{t=1}^T \nu^t)$ achieves $\mathrm{NEGap}(\overline{\mu}, \overline{\nu}) \le \varepsilon$ with probability at least $1 - \delta$, as long as*

$$T \ge \mathcal{O}(H^3(XA + YB)\iota/\varepsilon^2),$$

where $\iota := \log(10(XA + YB)/\delta)$ is a log factor. The total amount of episodes played is at most

$$2H \cdot T = \mathcal{O}(H^4(XA + YB)\iota/\varepsilon^2).$$

Corollary 7 shows that Balanced CFR requires $\widetilde{\mathcal{O}}(H^4(XA + YB)/\varepsilon^2)$ episodes for learning an ε-NE, which is H times larger than Balanced OMD but otherwise also near-optimal with respect to the lower bound (Theorem F.1) modulo an $\widetilde{\mathcal{O}}(\mathrm{poly}(H))$ factor. This improves significantly over the current best sample complexity achieved by CFR-type algorithms, which are either $\mathrm{poly}(X, Y, A, B)/\varepsilon^4$ (Farina & Sandholm, 2021), or potentially $\mathrm{poly}(X, Y, A, B)/\varepsilon^2$ using the MCCFR framework of (Lanctot et al., 2009; Farina et al., 2020c) but without any known such instantiation. The proof of Theorem 6 (and a brief overview) can be found in Appendix G.2.

In Appendix H we show that Balanced CFR can also be instantiated with Regret Matching (Hart & Mas-Colell, 2000) and enjoys similar sharp in X, Y sample complexities (though worse in A, B).

5 Extension to Multi-player general-sum games

Both Balanced OMD and Balanced CFR can be extended to learning Coarse Correlated Equilibria (Celli et al., 2019) in multi-player general-sum IIEFGs. Due to space constraint, we defer the definitions and the result to Appendix I and Theorem I.3.

6 Conclusion

This paper presents the first line of algorithms for learning an ε-NE in two-player zero-sum IIEFGs with near-optimal $\widetilde{\mathcal{O}}((XA + YB)/\varepsilon^2)$ sample complexity. We achieve this by new variants of both OMD and CFR type algorithms that incorporate suitable balanced exploration policies. We believe our work opens up many interesting future directions, such as empirical verification of our balanced algorithms, or how to learn IIEFGs with large state/action spaces efficiently using function approximation.

REFERENCES

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

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Noam Brown and Tuomas Sandholm. Regret transfer and parameter optimization. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Noam Brown and Tuomas Sandholm. Regret-based pruning in extensive-form games. In *NIPS*, pp. 1972–1980, 2015.

Noam Brown and Tuomas Sandholm. Reduced space and faster convergence in imperfect-information games via pruning. In *International conference on machine learning*, pp. 596–604. PMLR, 2017.

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

Noam Brown, Christian Kroer, and Tuomas Sandholm. Dynamic thresholding and pruning for regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International conference on machine learning*, pp. 793–802. PMLR, 2019.

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

Andrea Celli, Stefano Coniglio, and Nicola Gatti. Computing optimal ex ante correlated equilibria in two-player sequential games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 909–917, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33: 5527–5540, 2020.

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

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

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

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

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

Richard Gibson, Marc Lanctot, Neil Burch, Duane Szafron, and Michael Bowling. Generalized sampling and variance in counterfactual regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012a.

Richard G Gibson, Neil Burch, Marc Lanctot, and Duane Szafron. Efficient monte carlo counterfactual regret minimization in games with many player actions. In *NIPS*, pp. 1889–1897, 2012b.

Andrew Gilpin, Javier Pena, and Tuomas W Sandholm. First-order algorithm with o (ln (1/ε)) convergence for-equilibrium in two-person zero-sum games. 2008.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

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

Samid Hoda, Andrew Gilpin, Javier Pena, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512, 2010.

Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

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

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021a.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. *arXiv preprint arXiv:2106.03352*, 2021b.

Michael Johanson. Measuring the size of large no-limit poker games. *arXiv preprint arXiv:1302.7008*, 2013.

Michael Johanson, Nolan Bard, Marc Lanctot, Richard G Gibson, and Michael Bowling. Efficient nash equilibrium approximation through monte carlo counterfactual regret minimization. In *AAMAS*, pp. 837–846. Citeseer, 2012.

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

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

Vojtěch Kovařík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisỳ. Rethinking formal models of partially observable multiagent decision making. *arXiv preprint arXiv:1906.11110*, 2019.

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

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

Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Solving large sequential games with the excessive gap technique. *arXiv preprint arXiv:1810.03063*, 2018.

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

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

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

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

Viliam Lisỳ, Marc Lanctot, and Michael H Bowling. Online monte carlo counterfactual regret minimization for search in imperfect information games. In *AAMAS*, pp. 27–36, 2015.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pp. 1–22, 2022.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Remi Munos, Julien Perolat, Jean-Baptiste Lespiau, Mark Rowland, Bart De Vylder, Marc Lanctot, Finbarr Timbers, Daniel Hennes, Shayegan Omidshafiei, Audrunas Gruslys, et al. Fast computation of nash equilibria in imperfect information games. In *International Conference on Machine Learning*, pp. 7119–7129. PMLR, 2020.

John F Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49, 1950.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *arXiv preprint arXiv:1506.03271*, 2015.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Martin Schmid, Neil Burch, Marc Lanctot, Matej Moravcik, Rudolf Kadlec, and Michael Bowling. Variance reduction in monte carlo counterfactual regret minimization (vr-mccfr) for extensive form games using baselines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2157–2164, 2019.

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

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002. PMLR, 2020.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Oskari Tammelin. Solving large imperfect information games using cfr+. *arXiv preprint arXiv:1407.5042*, 2014.

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

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

Kevin Waugh, Dustin Morrill, James Andrew Bagnell, and Michael Bowling. Solving games with functional regret estimation. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. *arXiv preprint arXiv:1712.00579*, 2017.

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

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pp. 3674–3682. PMLR, 2020.

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

Kaiqing Zhang, Sham M Kakade, Tamer Başar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020.

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

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

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

A Related work

Computing NE from full feedback When the full game (transitions and rewards) is known, the problem of finding the NE is a min-max optimization problem over the two policies. Early works consider casting this min-max problem over the sequence-form policies as a linear program (Koller & Megiddo, 1992; Von Stengel, 1996; Koller et al., 1996). First-order algorithms are later proposed for solving the min-max problem directly, in particular by using proper regularizers such as the *dilated KL distance* (Gilpin et al., 2008; Hoda et al., 2010; Kroer et al., 2015; Lee et al., 2021). Another prevalent approach is Counterfactual Regret Minimization (CFR) (Zinkevich et al., 2007), which works by minimizing (local) *counterfactual regrets* at each infoset separately using any regret minimization algorithm over the probability simplex such as Regret Matching or Hedge (Tammelin, 2014; Burch et al., 2019; Zhou et al., 2020; Farina et al., 2020b). As each CFR iteration involves traversing the entire game tree which can be slow or memory-inefficient, techniques based on sampling or approximation have been proposed to address this, such as Monte-Carlo CFR (MCCFR) (Lanctot et al., 2009; Gibson et al., 2012b;a; Johanson et al., 2012; Lisỳ et al., 2015; Schmid et al., 2019), function approximation of counterfactual values (Waugh et al., 2015; Brown et al., 2019), and pruning (Brown & Sandholm, 2015; Brown et al., 2017; Brown & Sandholm, 2017).

Learning NE from bandit feedback The MCCFR framework (Lanctot et al., 2009) provides a first line of approaches for learning an ε-NE in IIEFGs from bandit feedback, by feeding in sample-based unbiased loss estimators to CFR algorithms. This framework is then generalized by Farina et al. (2020c) to any regret minimization algorithm (not necessarily CFR). They analyze the concentration between the true regret and the regret on loss estimators, and propose to sample with a "balanced strategy" (equivalent to a special case of our balanced exploration policy) to enable a small concentration term. However, they do not bound the regret on loss estimators or give an end-to-end sample complexity guarantee. Farina & Sandholm (2021) instantiate this framework to give a sample complexity guarantee of $\widetilde{\mathcal{O}}(\mathrm{poly}(X, Y, A, B)/\varepsilon^4)$, by using an exploration rule that favors larger sub-games (similar to our balanced exploration policy but defined through the number

of terminal *states* instead of infosets). Our Balanced CFR algorithm (Section 4) can be seen as an instantiation of this framework using a more general balanced exploration policy in both the sampling and the local regret minimization steps.

Another line of work considers sample-based variants of Online Mirror Descent (OMD) algorithms. Farina et al. (2021) provide an algorithm with $\widetilde{\mathcal{O}}((X^4A^3 + Y^4B^3)/\varepsilon^2)$ sample complexity[4] by OMD with an unbiased loss estimator and the dilated KL distance. Kozuno et al. (2021) propose the IXOMD algorithm that achieves $\widetilde{\mathcal{O}}((X^2A + Y^2B)/\varepsilon^2)$ sample complexity using an implicit-exploration loss estimator. Our Balanced OMD (Section 3) can be seen as a variant of the IXOMD algorithm by using a new variant of the dilated KL distance.

Finally, Zhou et al. (2019); Zhang & Sandholm (2021) propose model-based exploration approaches combined with planning on the estimated models. Specifically, Zhou et al. (2019) use posterior sampling to obtain an $\widetilde{\mathcal{O}}(SAB/\varepsilon^2)$ sample complexity under the Bayesian setting assuming a correct prior. Zhang & Sandholm (2021) achieve $\widetilde{\mathcal{O}}(S^2AB/\varepsilon^2)$ sample complexity by constructing confidence bounds for the transition model. Both sample complexities are polynomial in S (the number of underlying states) due to their need of estimating the full model, which could be much higher than $\mathrm{poly}(X, Y)$. A comparison between the above existing results and ours is given in Table 1.

Markov games without tree structure A related line of work considers learning equilibria in Markov Games (MGs) (Shapley, 1953) with perfect information, but without the tree structure assumed in IIEFGs. Sample-efficient algorithms for learning MGs from bandit feedback have been designed for learning NE in two-player zero-sum MGs either assuming access to a "simulator" or certain reachability assumptions, e.g. (Sidford et al., 2020; Zhang et al., 2020; Daskalakis et al., 2020; Wei et al., 2021) or in the exploration setting, e.g. (Wei et al., 2017; Bai & Jin, 2020; Xie et al., 2020; Bai et al., 2020; Liu et al., 2021; Chen et al., 2021; Jin et al., 2021b; Huang et al., 2021), as well as learning (Coarse) Correlated Equilibria in multi-player general-sum MGs, e.g. (Liu et al., 2021; Song et al., 2021; Jin et al., 2021a; Mao & Başar, 2022). As the settings of MGs in these work do not allow imperfect information, these results do not imply results for learning IIEFGs.

B Technical tools

The following Freedman's inequality can be found in (Agarwal et al., 2014, Lemma 9).

Lemma B.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*

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

C Bounds for regret minimizers

Here we collect regret bounds for various regret minimization algorithms on the probability simplex. For any algorithm that plays policy p_t in the t-th round and observes loss vector $\{\ell_t(a)\}_{a\in[A]} \in \mathbb{R}^A_{\geq 0}$, define its regret as

$$\mathrm{Regret}(T) := \max_{p^\star \in \Delta([A])} \sum_{t=1}^T \left\langle p_t, \widetilde{\ell}_t \right\rangle - \left\langle p^\star, \widetilde{\ell}_t \right\rangle.$$

C.1 Hedge

The following regret bound for Hedge is standard, see, e.g. (Lattimore & Szepesvári, 2020, Proposition 28.7).

[4]By plugging in an $\widetilde{\mathcal{O}}(X)$ upper bound for the dilated KL distance and optimizing the regret bound by setting $\eta = 1/\sqrt{X^2A^3T}$ in their Theorem 3.

Algorithm 3 Regret Minimization with Hedge (HEDGE)

Require: Learning rate $\eta > 0$.

1: Initialize $p_1(a) \leftarrow 1/A$ for all $a \in [A]$.
2: **for** iteration $t = 1, \ldots, T$ **do**
3: Receive loss vector $\left\{\widetilde{\ell}_t(a)\right\}_{a\in[A]}$.
4: Update action distribution via mirror descent:

$$p_{t+1}(a) \propto_a p_t(a) \exp\left(-\eta \widetilde{\ell}_t(a)\right).$$

Lemma C.1 (Regret bound for Hedge)**.** *Algorithm 3 with learning rate $\eta > 0$ achieves regret bound*

$$\text{Regret}(T) \leq \frac{\log A}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{a\in[A]} p_t(a) \widetilde{\ell}_t(a)^2.$$

C.2 REGRET MATCHING

Algorithm 4 Regret Minimization with Regret Matching (REGRETMATCHING)

1: Initialize $p_1(a) \leftarrow 1/A$ and $R_0(a) \leftarrow 0$ for all $a \in [A]$.
2: **for** iteration $t = 1, \ldots, T$ **do**
3: Receive loss vector $\left\{\widetilde{\ell}_t(a)\right\}_{a\in[A]}$.
4: Update instantaneous regret and cumulative regret for all $a \in [A]$:

$$r_t(a) \leftarrow \left\langle p_t, \widetilde{\ell}_t \right\rangle - \widetilde{\ell}_t(a) \quad \text{and} \quad R_t(a) \leftarrow R_{t-1}(a) + r_t(a).$$

5: Compute action distribution by regret matching:

$$p_{t+1}(a) \leftarrow \frac{[R_t(a)]_+}{\sum_{a'\in[A]} [R_t(a')]_+} = \frac{\left[\sum_{t=1}^{T} \left\langle p_t, \tilde{\ell}_t \right\rangle - \widetilde{\ell}_t(a)\right]_+}{\sum_{a'\in[A]} \left[\sum_{t=1}^{T} \left\langle p_t, \tilde{\ell}_t \right\rangle - \widetilde{\ell}_t(a')\right]_+}.$$

In the edge case where $[R_t(a)]_+ = 0$ for all $a \in [A]$, set $p_{t+1}(a) \leftarrow 1/A$ to be the uniform distribution.

The following regret bound for Regret Matching is standard, see, e.g. (Cesa-Bianchi & Lugosi, 2006; Brown & Sandholm, 2014). For completeness, here we provide a proof along with an alternative form of bound useful for our purpose (Remark C.3). Note that here η is not the learning rate but rather an arbitrary positive value (i.e. the right-hand side is an upper bound on the regret for any $\eta > 0$). Algorithm 4 itself does not require any learning rate.

Lemma C.2 (Regret bound for Regret Matching)**.** *Algorithm 4 achieves the following regret bound for* any $\eta > 0$*:*

$$\text{Regret}(T) \leq \Big[\sum_{t=1}^{T} \sum_{a\in[A]} \Big(\left\langle p_t, \widetilde{\ell}_t \right\rangle - \widetilde{\ell}_t(a)\Big)^2\Big]^{1/2} \leq \frac{1}{\eta} + \frac{\eta}{4} \sum_{t=1}^{T} \sum_{a\in[A]} \Big(\left\langle p_t, \widetilde{\ell}_t \right\rangle - \widetilde{\ell}_t(a)\Big)^2.$$

Proof. By the fact that $(a+b)_+^2 \leq a_+^2 + 2a_+b + b^2$, we have

$$[R_t(a)]_+^2 \leq [R_{t-1}(a)]_+^2 + 2[R_{t-1}(a)]_+ r_t(a) + r_t(a)^2. \tag{14}$$

Then by the definition of $p_t(a)$ and $r_t(a)$, we have

$$\begin{aligned} \sum_{a\in[A]} [R_{t-1}(a)]_+ r_t(a) &= \sum_{a\in[A]} [R_{t-1}(a)]_+ \Big(\sum_{a'\in[A]} p_t(a') \widetilde{\ell}_t(a') - \widetilde{\ell}_t(a)\Big) \\ &= \sum_{a\in[A]} [R_{t-1}(a)]_+ \widetilde{\ell}_t(a) - \sum_{a\in[A]} [R_{t-1}(a)]_+ \widetilde{\ell}_t(a) = 0. \end{aligned} \tag{15}$$

Then summing over a in Eq. (14) and using Eq. (15), we get

$$\sum_{a\in[A]}[R_T(a)]_+^2 \le \sum_{a\in[A]}[R_{T-1}(a)]_+^2 + 2\sum_{a\in[A]}[R_{T-1}(a)]_+ r_T(a) + \sum_{a\in[A]} r_T(a)^2$$
$$= \sum_{a\in[A]}[R_{T-1}(a)]_+^2 + \sum_{a\in[A]} r_T(a)^2 \le \sum_{t=1}^{T}\sum_{a\in[A]} r_t(a)^2.$$

Using that $\max_a R_T(a) \le \max_a [R_T(a)]_+ \le (\sum_{a\in[A]}[R_T(a)]_+^2)^{1/2}$ gives the regret bound

$$\mathrm{Regret}(T) = \max_{a\in[A]} R_T(a) \le \left(\sum_{t=1}^{T}\sum_{a\in[A]} r_t(a)^2\right)^{1/2} = \left(\sum_{t=1}^{T}\sum_{a\in[A]}\left(\left\langle p_t, \widetilde{\ell}_t\right\rangle - \widetilde{\ell}_t(a)\right)^2\right)^{1/2}.$$

The claimed bound with η follows directly from the inequality $\sqrt{z} \le 1/\eta + \eta z/4$ for any $\eta > 0$, $z \ge 0$. □

Remark C.3. *The quantity* $\sum_{a\in[A]}\left(\left\langle p_t, \widetilde{\ell}_t\right\rangle - \widetilde{\ell}_t(a)\right)^2$ *above can be upper bounded as*

$$\sum_{a\in[A]}\left(\left\langle p_t, \widetilde{\ell}_t\right\rangle - \widetilde{\ell}_t(a)\right)^2 \le \sum_{a\in[A]}\left(\left\langle p_t, \widetilde{\ell}_t\right\rangle^2 + \widetilde{\ell}_t(a)^2\right)$$
$$= A\left\langle p_t, \widetilde{\ell}_t\right\rangle^2 + \left\|\widetilde{\ell}_t\right\|_2^2 \le A\sum_{a\in[A]}\left(p_t(a)\widetilde{\ell}_t(a)^2 + (1/A)\widetilde{\ell}_t(a)^2\right)$$
$$= 2A\sum_{a\in[A]}\bar{p}_t(a)\widetilde{\ell}_t(a)^2,$$

where $\bar{p}_t(a) = [p_t(a) + (1/A)]/2$ *is a probability distribution over* $[A]$.

As a consequence, we get an upper bound on the regret of Regret Matching algorithm by

$$\mathrm{Regret}(T) \le \frac{1}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{a\in[A]}(A\bar{p}_t(a))\widetilde{\ell}_t(a)^2.$$

Comparing to the bound of Hedge (Lemma C.1), the above regret bound for Regret Matching has a similar form except for replacing $\log A$ *by* 1 *and replacing* p_t *by* $A\bar{p}_t$.

D Properties of the game

D.1 Basic properties

For any opponent (min-player) policy $\nu \in \Pi_{\min}$, define

$$p^{\nu}_{1:h}(x_h) := \sum_{s_h\in x_h} p_{1:h}(s_h)\nu_{1:h-1}(y(s_{h-1}), b_{h-1}) \quad \text{for all } h\in[H],\ x_h\in\mathcal{X}_h.$$

Intuitively, $p^{\nu}_{1:h}(x_h)$ measures the environment and the opponent's contribution in the reaching probability of x_h.

Lemma D.1 (Properties of $p^{\nu}_{1:h}(x_h)$)**.** *The following holds for any* $\nu \in \Pi_{\min}$*:*

(a) For any policy $\mu \in \Pi_{\max}$*, we have*

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

(b) $0 \le p^{\nu}_{1:h}(x_h) \le 1$ *for all* h, x_h.

Proof. For (a), notice that

$$\mu_{1:h}(x_h, a_h)p^{\nu}_{1:h}(x_h) = \sum_{s_h \in x_h} p_{1:h}(s_h) \cdot \mu_{1:h}(x_h, a_h) \cdot \nu_{1:h-1}(y(s_{h-1}), b_{h-1})$$
$$= \sum_{s_h \in x_h} \mathbb{P}^{\mu,\nu}(\text{visit } (s_h, a_h)) = \mathbb{P}^{\mu,\nu}(\text{visit } (x_h, a_h)).$$

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

For (b), fix any $x_h \in \mathcal{X}_h$. Clearly $p^{\nu}_{1:h}(x_h) \geq 0$. Choose any $a_h \in \mathcal{A}$, and choose policy $\mu^{x_h,a_h} \in \Pi_{\max}$ such that $\mu^{x_h,a_h}_{1:h}(x_h, a_h) = 1$ (such μ^{x_h,a_h} exists, for example, by deterministically taking all actions prescribed in infoset x_h at all ancestors of x_h). For this μ^{x_h,a_h}, using (a), we have

$$p^{\nu}_{1:h}(x_h) = \mu^{x_h,a_h}_{1:h}(x_h, a_h) \cdot p^{\nu}_{1:h}(x_h) \leq \sum_{(x'_h,a'_h) \in \mathcal{X}_h \times \mathcal{A}} \mu^{x_h,a_h}_{1:h}(x'_h, a'_h) \cdot p^{\nu}_{1:h}(x'_h) = 1.$$

This shows part (b). □

Corollary D.2. *For any policy $\mu \in \Pi_{\max}$ and $h \in [H]$, we have*

$$\sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h)\ell^t_h(x_h, a_h) \leq 1.$$

Proof. Notice by definition

$$\ell^t_h(x_h, a_h) = \sum_{s_h \in x_h, b_h \in \mathcal{B}_h} p_{1:h}(s_h)\nu^t_{1:h}(y(s_h), b_h)(1 - r_h(s_h, a_h, b_h)) \leq p^{\nu}_{1:h}(x_h),$$

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

Lemma D.3. *For any $h \in [H]$, the counterfactual loss function L^t_h defined in (10) satisfies the bound*

(a) For any policy $\mu \in \Pi_{\max}$, we have

$$\sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h)L^t_h(x_h, a_h) \leq H - h + 1.$$

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

$$0 \leq L^t_h(x_h, a_h) \leq p^{\nu^t}_{1:h}(x_h) \cdot (H - h + 1).$$

Proof. Part (a) follows from the fact that

$$\sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h)L^t_h(x_h, a_h) = \mathbb{E}_{\mu,\nu^t}\left[\sum_{h'=h}^{H} r_{h'}\right] \leq H - h + 1,$$

where the first equality follows from the definition of the loss functions ℓ_h and L_h in (3), (10).

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

$$L^t_h(x_h, a_h) = \mu^{x_h,a_h}_{1:h}(x_h, a_h)L^t_h(x_h, a_h) = \mathbb{E}_{\mu^{x_h,a_h},\nu^t}\left[\mathbf{1}\{\text{visit } x_h, a_h\} \cdot \sum_{h'=h}^{H} r_{h'}\right]$$
$$= \mathbb{P}_{\mu^{x_h,a_h},\nu^t}(\text{visit } x_h, a_h) \cdot \mathbb{E}_{\mu^{x_h,a_h},\nu^t}\left[\sum_{h'=h}^{H} r_{h'} \middle| \text{visit } x_h, a_h\right]$$
$$\leq \mu^{x_h,a_h}_{1:h}(x_h, a_h)p^{\nu^t}_{1:h}(x_h) \cdot (H - h + 1) = p^{\nu^t}_{1:h}(x_h) \cdot (H - h + 1).$$

□

Definition of average policies For two-player zero-sum IIEFGs, we define the average policy of the max-player $\overline{\mu} = \frac{1}{T}\sum_{t=1}^{T}\mu^t$ (in conditional form) by

$$\overline{\mu}_h(a_h|x_h) := \frac{\sum_{t=1}^{T}\mu^t_{1:h}(x_h,a_h)}{\sum_{t=1}^{T}\mu^t_{1:h-1}(x_h)}, \tag{16}$$

for any h and $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$. It is straightforward to check that this $\overline{\mu}$ is exactly the averaging of μ^t in the sequence-form representation (see e.g. (Kozuno et al., 2021, Theorem 1)):

$$\overline{\mu}_{1:h}(x_h,a_h) = \frac{1}{T}\sum_{t=1}^{T}\mu^t_{1:h}(x_h,a_h) \quad \text{for all } (h, x_h, a_h). \tag{17}$$

Both expressions above can be used as the definition interchangably. The average policy of the min-player $\overline{\nu} = \frac{1}{T}\sum_{t=1}^{T}\nu^t$ is defined similarly.

D.2 Balanced exploration policy

Lemma D.4 (Balancing property of $\mu^{\star,h}$)**.** *For any max-player's policy $\mu \in \Pi_{\max}$ and any $h \in [H]$, we have*

$$\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}(x_h,a_h)}{\mu^{\star,h}_{1:h}(x_h,a_h)} = X_h A.$$

Lemma D.4 states that $\mu^{\star,h}$ is a good exploration policy in the sense that the distribution mismatch between it and *any* $\mu \in \Pi_{\max}$ has bounded L_1 norm. Further, the bound $X_h A$ is non-trivial—For example, if we replace $\mu^{\star,h}_{1:h}$ with the uniform policy $\mu^{\mathrm{unif}}_{1:h}(x_h,a_h) = 1/A^h$, the left-hand side can be as large as $X_h A^h$ in the worst case.

Proof of Lemma D.4 We have

$$\begin{aligned}
&\sum_{x_h,a_h}\frac{\mu_{1:h}(x_h,a_h)}{\mu^{\star,h}_{1:h}(x_h,a_h)} \\
&= \sum_{x_{h-1},a_{h-1}}\sum_{(x_h,a_h)\in\mathcal{C}(x_{h-1},a_{h-1})\times\mathcal{A}} \frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})\cdot\mu_h(a_h|x_h)}{\mu^{\star,h}_{1:(h-1)}(x_{h-1},a_{h-1})\cdot(1/A)} \\
&\overset{(i)}{=} A\cdot\sum_{x_{h-1},a_{h-1}}\sum_{x_h\in\mathcal{C}(x_{h-1},a_{h-1})}\frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})}{\mu^{\star,h}_{1:(h-1)}(x_{h-1},a_{h-1})} \\
&= A\cdot\sum_{x_{h-1},a_{h-1}}\frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})}{\mu^{\star,h}_{1:(h-1)}(x_{h-1},a_{h-1})}\cdot|\mathcal{C}_h(x_{h-1},a_{h-1})| \\
&\overset{(ii)}{=} A\cdot\sum_{x_{h-2},a_{h-2}}\sum_{(x_{h-1},a_{h-1})\in\mathcal{C}(x_{h-2},a_{h-2})\times\mathcal{A}}\frac{\mu_{1:(h-2)}(x_{h-2},a_{h-2})\mu_{h-1}(a_{h-1}|x_{h-1})}{\mu^{\star,h}_{1:(h-2)}(x_{h-2},a_{h-2})\cdot|\mathcal{C}_h(x_{h-1},a_{h-1})|/|\mathcal{C}_h(x_{h-1})|}\cdot|\mathcal{C}_h(x_{h-1},a_{h-1})| \\
&= A\cdot\sum_{x_{h-2},a_{h-2}}\sum_{(x_{h-1},a_{h-1})\in\mathcal{C}(x_{h-2},a_{h-2})\times\mathcal{A}}\frac{\mu_{1:(h-2)}(x_{h-2},a_{h-2})\mu_{h-1}(a_{h-1}|x_{h-1})}{\mu^{\star,h}_{1:(h-2)}(x_{h-2},a_{h-2})}\cdot|\mathcal{C}_h(x_{h-1})| \\
&= A\cdot\sum_{x_{h-2},a_{h-2}}\sum_{(x_{h-1},a_{h-1})\in\mathcal{C}(x_{h-2},a_{h-2})\times\mathcal{A}}\frac{\mu_{1:(h-2)}(x_{h-2},a_{h-2})\mu_{h-1}(a_{h-1}|x_{h-1})}{\mu^{\star,h}_{1:(h-2)}(x_{h-2},a_{h-2})}\cdot|\mathcal{C}_h(x_{h-1})| \\
&\overset{(iii)}{=} A\cdot\sum_{x_{h-2},a_{h-2}}\frac{\mu_{1:(h-2)}(x_{h-2},a_{h-2})}{\mu^{\star,h}_{1:(h-2)}(x_{h-2},a_{h-2})}\cdot|\mathcal{C}_h(x_{h-2},a_{h-2})| \\
&= \dots \\
&= A\cdot\sum_{x_1,a_1}\frac{\mu_1(a_1|x_1)}{|\mathcal{C}_h(x_1,a_1)|/|\mathcal{C}_h(x_1)|}\cdot|\mathcal{C}_h(x_1,a_1)| \\
&= A\cdot\sum_{x_1,a_1}\mu_1(a_1|x_1)\cdot|\mathcal{C}_h(x_1)|
\end{aligned}$$

$$= A \cdot \sum_{x_1} |\mathcal{C}_h(x_1)| = A \cdot |\mathcal{C}_h(\emptyset)| = X_h A.$$

Above, (i) used the definition of $\mu_h^{\star,h}$ and the fact that $\sum_{a_h \in \mathcal{A}} \mu_h(a_h|x_h) = 1$ for any μ, x_h; (ii) used the definition of $\mu_{h-1}^{\star,h}$; (iii) used the fact that $\sum_{x_{h-1} \in \mathcal{C}(x_{h-2}, a_{h-2})} |\mathcal{C}_h(x_{h-1})| = |\mathcal{C}_h(x_{h-2}, a_{h-2})|$ which follows by the additivity of the number of descendants; and the rest followed by performing the same operations repeatedly. □

The following corollary is similar to the lower bound in (Farina et al., 2020c, Appendix A.3).

Corollary D.5. *We have*

$$\mu_{1:h}^{\star,h}(x_h, a_h) \geq \frac{1}{X_h A}$$

for any $h \in [H]$ and $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$.

Proof. Choose some deterministic policy μ s.t. $\mu_{1:h}(x_h, a_h) = 1$ in Lemma D.4 and noticing each term in the summation is non-negative,

$$\frac{\mu_{1:h}(x_h, a_h)}{\mu_{1:h}^{\star,h}(x_h, a_h)} \leq X_h A.$$

□

D.2.1 Interpretation as a transition probability

We now provide an intepretation of the balanced exploration policy $\mu_{1:h}^{\star,h}$: its inverse $1/\mu_{1:h}^{\star,h}$ can be viewed as the (product) of a "transition probability" over the game tree for the max player. As a consequence, this interpretation also provides an alternative proof of Lemma D.4.

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

$$p_{1:h}^{\star,h}(x_h) = \frac{|\mathcal{C}_h(x_1)|}{X_h} \prod_{k=1}^{h-1} p_k^{\star,h}(x_{k+1}|x_k, a_k) = \frac{|\mathcal{C}_h(x_1)|}{X_h} \prod_{k=1}^{h-1} \frac{|\mathcal{C}_h(x_{k+1})|}{|\mathcal{C}_h(x_k, a_k)|}. \tag{18}$$

Lemma D.6. *For any $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$, the sequence form of the transition $p_{1:h}^{\star,h}(x_h)$ and the sequence form of balanced exploration strategy $\mu_{1:h}^{\star,h}(x_h, a_h)$ are related by*

$$p_{1:h}^{\star,h}(x_h) = \frac{1}{X_h A \cdot \mu_{1:h}^{\star,h}(x_h, a_h)}. \tag{19}$$

Furthermore, for any max player's policy $\mu \in \Pi_{\max}$ and any $h \in [H]$, we have

$$\sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h) p_{1:h}^{\star,h}(x_h) = 1. \tag{20}$$

Proof of Lemma D.6 By the definition of the balanced transition probability as in Eq. (18) and the balanced exploration strategy as in Eq. (5), we have

$$\frac{1}{X_h A \cdot \mu_{1:h}^{\star,h}(x_h, a_h)} = \frac{1}{X_h A} \prod_{k=1}^{h-1} \frac{|\mathcal{C}_h(x_k)|}{|\mathcal{C}_h(x_k, a_k)|} \times A = \frac{|\mathcal{C}_h(x_1)|}{X_h} \prod_{k=1}^{h-1} \frac{|\mathcal{C}_h(x_{k+1})|}{|\mathcal{C}_h(x_k, a_k)|} = p_{1:h}^{\star,h}(x_h).$$

where the second equality used the property that $|\mathcal{C}_h(x_h)| = 1$. This proves Eq. (19). The proof of Eq. (20) is similar to the proof of Lemma D.1 (a). □

Alternative proof of Lemma D.4 Lemma D.4 follows as a direct consequence of Eq. (19) and (20) in Lemma D.6. □

D.3 BALANCED DILATED KL

Lemma D.7 (Bound on balanced dilated KL)**.** *Let* $\mu^{\mathrm{unif}} \in \Pi_{\max}$ *denote the uniform policy:* $\mu_h^{\mathrm{unif}}(a_h|x_h) = 1/A$ *for all* (h, x_h, a_h)*. Then we have*

$$\max_{\mu^\dagger \in \Pi_{\max}} \mathrm{D}^{\mathrm{bal}}(\mu^\dagger \| \mu^{\mathrm{unif}}) \leq XA \log A.$$

Proof. We have

$$\begin{aligned}
\max_{\mu^\dagger \in \Pi_{\max}} \mathrm{D}^{\mathrm{bal}}(\mu^\dagger \| \mu^{\mathrm{unif}}) &= \max_{\mu^\dagger \in \Pi_{\max}} \sum_{h=1}^{H} \sum_{x_h, a_h} \frac{\mu^\dagger_{1:h}(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \log \frac{\mu^\dagger_h(a_h|x_h)}{\mu^{\mathrm{unif}}_h(a_h|x_h)} \\
&= \max_{\mu^\dagger \in \Pi_{\max}} \sum_{h=1}^{H} \sum_{x_h, a_h} \frac{\mu^\dagger_{1:h}(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \Big(\log \mu^\dagger_h(a_h|x_h) + \log A \Big) \\
&\overset{(i)}{\leq} \log A \sum_{h=1}^{H} \max_{\mu^\dagger \in \Pi_{\max}} \sum_{x_h, a_h} \frac{\mu^\dagger_{1:h}(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \\
&\overset{(ii)}{=} \log A \sum_{h=1}^{H} X_h A = XA \log A,
\end{aligned}$$

where (i) is because $\mu^\dagger_h(a_h|x_h) \log \mu^\dagger_h(a_h|x_h) \leq 0$ (recalling that each sequence form $\mu^\dagger_{1:h}(x_h, a_h)$ contains the term $\mu^\dagger_h(a_h|x_h)$), and (ii) uses the balancing property of $\mu^{\star,h}$ (Lemma D.4). □

D.3.1 INTERPRETATION OF BALANCED DILATED KL

We present an interpretation of the balanced dilated KL (6) as a KL distance between the reaching probabilities under the "balanced transition" (18) on the max player's game tree.

For any policy $\mu \in \Pi_{\max}$, we define its *balanced transition reaching probability* $\mathbb{P}_h^{\mu,\star}(x_h, a_h)$ as

$$\mathbb{P}_h^{\mu,\star}(x_h, a_h) = \mu_{1:h}(x_h, a_h) p^{\star,h}_{1:h}(x_h). \tag{21}$$

This is a probability measure on $\mathcal{X}_h \times \mathcal{A}$ ensured by Lemma D.6. For any two probability distribution p and q, we denote $\mathrm{KL}(p\|q)$ to be their KL divergence.

Lemma D.8. *For any tuple of max-player's policies* $\mu, \nu \in \Pi_{\max}$*, we have*

$$\mathrm{D}^{\mathrm{bal}}(\mu\|\nu) = \sum_{h=1}^{H} (X_h A) \mathrm{KL}(\mathbb{P}_h^{\mu_{1:h},\star} \| \mathbb{P}_h^{\mu_{1:h-1}\nu_h,\star}). \tag{22}$$

Proof of Lemma D.8 By Eq. (21) and by the definition of KL divergence, we have

$$\begin{aligned}
&(X_h A) \mathrm{D}^{\mathrm{kl}}(\mathbb{P}_h^{\mu_{1:h},\star} \| \mathbb{P}_h^{\mu_{1:h-1}\nu_h,\star}) \\
&= (X_h A) \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h) p^{\star,h}_{1:h}(x_h) \log \Big[\frac{\mu_{1:h}(x_h, a_h) p^{\star,h}_{1:h}(x_h)}{\mu_{1:h-1}(x_{h-1}, a_{h-1}) \nu_h(x_h|a_h) p^{\star,h}_{1:h}(x_h)} \Big] \\
&= \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{\mu_{1:h}(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \log \Big[\frac{\mu_h(a_h|x_h)}{\nu_h(a_h|x_h)} \Big],
\end{aligned} \tag{23}$$

where the last equality is by Lemma D.6. Comparing with the definition of $\mathrm{D}^{\mathrm{bal}}$ as in Eq. (6) concludes the proof. □

Algorithm 5 Implementation of Balanced OMD update

Require: Current policy μ^t; Trajectory $(x_1^t, a_1^t, \ldots, x_H^t, a_H^t)$; learning rate $\eta > 0$;
Loss vector $\left\{\widetilde{\ell}_h^t(x_h, a_h)\right\}_{h,x_h,a_h}$ that is non-zero only on $(x_h, a_h) = (x_h^t, a_h^t)$.

1: Set $Z_{H+1}^t \leftarrow 1$.
2: **for** $h = H, \ldots, 1$ **do**
3: Compute normalization constant

$$Z_h^t \leftarrow 1 - \mu_h^t(a_h^t|x_h^t) + \mu_h^t(a_h^t|x_h^t) \cdot \exp\left(-\eta\mu_{1:h}^{\star,h}(x_h^t, a_h^t)\widetilde{\ell}_h^t(x_h^t, a_h^t) + \frac{\mu_{1:h}^{\star,h}(x_h^t, a_h^t)\log Z_{h+1}^t}{\mu_{1:h+1}^{\star,h+1}(x_{h+1}^t, a_{h+1}^t)}\right).$$

4: Update policy at x_h^t:

$$\mu_h^{t+1}(a_h|x_h^t) \leftarrow \begin{cases} \mu_h^t(a_h|x_h^t) \cdot \exp\left(-\eta\mu_{1:h}^{\star,h}(x_h^t, a_h^t)\widetilde{\ell}_h^t(x_h^t, a_h^t) + \frac{\mu_{1:h}^{\star,h}(x_h^t, a_h^t)\log Z_{h+1}^t}{\mu_{1:h+1}^{\star,h+1}(x_{h+1}^t, a_{h+1}^t)} - \log Z_h^t\right) & \text{if } a_h = a_h^t, \\ \mu_h^t(a_h|x_h^t) \cdot \exp(-\log Z_h^t) & \text{otherwise.} \end{cases}$$

5: Set $\mu_h^{t+1}(\cdot|x_h) \leftarrow \mu_h^t(\cdot|x_h)$ for all $x_h \in \mathcal{X}_h \setminus \{x_h^t\}$.

Ensure: Updated policy μ^{t+1}.

E Proofs for Section 3

E.1 Efficient implementation for Update (9)

Lemma E.1. *Algorithm 5 indeed solves the optimization problem (9):*

$$\mu^{t+1} \leftarrow \arg\min_{\mu \in \Pi_{\max}} \left\langle \mu, \widetilde{\ell}^t \right\rangle + \frac{1}{\eta}\mathrm{D}^{\mathrm{bal}}(\mu\|\mu^t).$$

Proof. First, by the sparsity of the loss estimator $\widetilde{\ell}^t$ (cf. (8)), the above objective can be written succinctly as

$$\left\langle \mu, \widetilde{\ell}^t \right\rangle + \frac{1}{\eta}\mathrm{D}^{\mathrm{bal}}(\mu\|\mu^t) \tag{24}$$

$$= \sum_{h=1}^{H}\sum_{x_h,a_h}\mu_{1:h}(x_h, a_h)\left[\widetilde{\ell}_h^t(x_h, a_h) + \frac{1}{\eta\mu_{1:h}^{\star,h}(x_h, a_h)}\log\frac{\mu_h(a_h|x_h)}{\mu_h^t(a_h|x_h)}\right]$$

$$= \sum_{h=1}^{H}\sum_{x_h}\mu_{1:h-1}(x_h)\left[\left\langle \mu_h(\cdot|x_h), \widetilde{\ell}_h^t(x_h, \cdot)\right\rangle + \frac{\mathrm{KL}\left(\mu_h(\cdot|x_h)||\mu_h^t(\cdot|x_h)\right)}{\eta\mu_{1:h}^{\star,h}(x_h, a_h)}\right]$$

$$= \sum_{h=1}^{H}\left\{\mu_{1:h-1}(x_h^t)\left[\mu_h(a_h^t|x_h^t)\widetilde{\ell}_h^t(x_h^t, a_h^t) + \frac{\mathrm{KL}\left(\mu_h(\cdot|x_h)||\mu_h^t(\cdot|x_h)\right)}{\eta\mu_{1:h}^{\star,h}(x_h^t, a_h)}\right] + \sum_{x_h \neq x_h^t}\mu_{1:h-1}(x_h)\frac{\mathrm{KL}\left(\mu_h(\cdot|x_h)||\mu_h^t(\cdot|x_h)\right)}{\eta\mu_{1:h}^{\star,h}(x_h, a_h)}\right\}. \tag{25}$$

We now show the equivalence by backward induction over $h = H, \ldots, 1$. For $h = H$, we can optimize over the H-th layer directly to see

$$\begin{aligned} \mu_H^{t+1}(a_H|x_H^t) &\propto_{a_H} \mu_H^t(a_H|x_H^t)\exp\left\{-\eta\mu_{1:h}^{\star,h}(x_h^t, a_h)\widetilde{\ell}_H^t(x_H^t, a_H)\right\} \\ &= \mu_H^t(a_H|x_H^t)\exp\left\{-\eta\widetilde{\ell}_H^t(x_H^t, a_H) - \log Z_H^t\right\}, \end{aligned}$$

where $Z_H^t > 0$ is the normalization constant. For all non-visited $x_H \neq x_H^t$, by equation (25) and non-negativity of KL divergence, the object must be minimized at $\mu_H^{t+1}(\cdot|x_H) = \mu_h^t(\cdot|x_H)$.

If the claim holds from layer $h+1$ to H, consider the h-th layer. Plug in the proved optimizer after layer h, the objective (25) can be written as

$$\sum_{h'=1}^{H}\sum_{x_{h'},a_{h'}}\mu_{1:h'}(x_{h'}, a_{h'})\left[\widetilde{\ell}_{h'}^t(x_{h'}, a_{h'}) + \frac{1}{\eta\mu_{1:h'}^{\star,h'}(x_{h'}, a_{h'})}\log\frac{\mu_{h'}(a_{h'}|x_{h'})}{\mu_{h'}^t(a_{h'}|x_{h'})}\right]$$

$$
\begin{aligned}
&= \sum_{h'=1}^{H} \sum_{x_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \widetilde{\ell}^t_{h'}(x_{h'}, \cdot) \right\rangle + \frac{\mathrm{KL}\left(\mu_{h'}(\cdot|x_{h'}) || \mu^t_{h'}(\cdot|x_{h'})\right)}{\eta \mu^{\star,h'}_{1:h'}(x_{h'}, a_{h'})} \right] \\
&= \sum_{h'=1}^{h} \sum_{x_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \widetilde{\ell}^t_{h'}(x_{h'}, \cdot) \right\rangle + \frac{\mathrm{KL}\left(\mu_{h'}(\cdot|x_{h'}) || \mu^t_{h'}(\cdot|x_{h'})\right)}{\eta \mu^{\star,h'}_{1:h'}(x_{h'}, a_{h'})} \right] \\
&\quad + \sum_{h'=h+1}^{H} \left[\frac{\mu_{1:h'}(x^t_{h'}, a^t_{h'}) \log Z^t_{h'+1}}{\eta \mu^{\star,h'+1}_{1:h'+1}(x^t_{h'+1}, a^t_{h'+1})} - \frac{\mu_{1:h'-1}(x^t_{h'-1}, a^t_{h'-1}) \log Z^t_{h'}}{\eta \mu^{\star,h'}_{1:h'}(x^t_{h'}, a^t_{h'})} \right] \\
&= \sum_{h'=1}^{h} \sum_{x_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \widetilde{\ell}^t_{h'}(x_{h'}, \cdot) \right\rangle + \frac{\mathrm{KL}\left(\mu_{h'}(\cdot|x_{h'}) || \mu^t_{h'}(\cdot|x_{h'})\right)}{\eta \mu^{\star,h'}_{1:h'}(x_{h'}, a_{h'})} \right] - \frac{\mu_{1:h}(x^t_h, a^t_h) \log Z^t_{h+1}}{\eta \mu^{\star,h+1}_{1:h+1}(x^t_{h+1}, a^t_{h+1})} \\
&= \sum_{h'=1}^{h-1} \sum_{x_{h'}} \mu_{1:h'-1}(x_{h'}) \left[\left\langle \mu_{h'}(\cdot|x_{h'}), \widetilde{\ell}^t_{h'}(x_{h'}, \cdot) \right\rangle + \frac{\mathrm{KL}\left(\mu_{h'}(\cdot|x_{h'}) || \mu^t_{h'}(\cdot|x_{h'})\right)}{\eta \mu^{\star,h'}_{1:h'}(x_{h'}, a_{h'})} \right] \\
&\quad + \mu_{1:h-1}(x^t_h) \left[\mu_h(a^t_h|x^t_h) \Big(\widetilde{\ell}^t_h(x^t_h, a^t_h) - \frac{\log Z^t_{h+1}}{\eta \mu^{\star,h+1}_{1:h+1}(x^t_{h+1}, a^t_{h+1})} \Big) + \frac{\mathrm{KL}\left(\mu_h(\cdot|x^t_h) || \mu^t_h(\cdot|x^t_h)\right)}{\eta \mu^{\star,h}_{1:h}(x^t_h, a_h)} \right] \\
&\quad + \sum_{x_h \neq x^t_h} \mu_{1:h-1}(x_h) \frac{\mathrm{KL}\left(\mu_h(\cdot|x_h) || \mu^t_h(\cdot|x_h)\right)}{\eta \mu^{\star,h}_{1:h}(x_h, a_h)}.
\end{aligned}
$$

Thus in the h layer we can optimize by setting

$$
\mu^{t+1}_h(a_h|x^t_h) = \mu^t_h(a_h|x^t_h) \exp\left\{ - \left[\eta \mu^{\star,h}_{1:h}(x^t_h, a_h) \widetilde{\ell}^t_h(x^t_h, a_h) - \frac{\mu^{\star,h}_{1:h}(x^t_h, a_h)}{\mu^{\star,h+1}_{1:h+1}(x^t_{h+1}, a^t_{h+1})} \log Z^t_{h+1} \right] \mathbf{1}\left\{a_h = a^t_h\right\} - \log Z^t_h \right\}.
$$

For all non-visited $x_h \neq x^t_h$, by non-negativity of KL divergence, the object must be minimized at $\mu^{t+1}_h(\cdot|x_h) = \mu^t_h(\cdot|x_h)$. This is exactly the update rule in Algorithm 5. □

E.2 Proof of Theorem 4

Overview of proof Our proof follows the usual analysis of OMD algorithms where the key is to bound a distance term and an algorithm-specific "stability" like term (cf. Lemma E.4 and its proof). Compared with existing OMD algorithms using the original dilated KL (Kozuno et al., 2021), our balanced dilated KL creates a "balancing effect" that preserves the distance term (Lemma D.7) and shaves off an X factor in the stability term (Lemma E.11 & E.12), which combine to yield a $\sqrt{X}$ improvement in the final regret bound.

Decompose the regret as

$$
\mathfrak{R}^T = \max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \left\langle \mu^t - \mu^\dagger, \ell^t \right\rangle \tag{26}
$$

$$
\leq \underbrace{\sum_{t=1}^{T} \left\langle \mu^t, \ell^t - \widetilde{\ell}^t \right\rangle}_{\text{BIAS}^1} + \underbrace{\max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \left\langle \mu^\dagger, \widetilde{\ell}^t - \ell^t \right\rangle}_{\text{BIAS}^2} + \underbrace{\max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \left\langle \mu^t - \mu^\dagger, \widetilde{\ell}^t \right\rangle}_{\text{REGRET}}. \tag{27}
$$

We now state three lemmas that bound each of the three terms above. Their proofs are presented in Section E.4, E.5, and E.6 respectively. Below, $\iota := \log(3HXA/\delta)$ denotes a log factor.

Lemma E.2 (Bound on BIAS^1)**.** *With probability at least* $1 - \delta/3$*, we have*

$$
\mathrm{BIAS}^1 \leq H\sqrt{2T\iota} + \gamma HT.
$$

Lemma E.3 (Bound on BIAS^2)**.** *With probability at least* $1 - \delta/3$*, we have*

$$
\mathrm{BIAS}^2 \leq XA\iota/\gamma.
$$

Lemma E.4 (Bound on REGRET)**.** *With probability at least* $1-\delta/3$*, we have*

$$\text{REGRET} \leq \frac{XA\log A}{\eta} + \eta H^3 T + \frac{\eta H^2 XA\iota}{\gamma}.$$

Putting the bounds together, we have that with probability at least $1-\delta$,

$$\mathfrak{R}^T \leq \frac{XA\log A}{\eta} + \eta H^3 T + \frac{\eta H^2 XA\iota}{\gamma} + H\sqrt{2T\iota} + \gamma HT + \frac{XA\iota}{\gamma}.$$

Set $\eta = \sqrt{\frac{XA\log A}{H^3 T}}$ and $\gamma = \sqrt{\frac{XA\iota}{TH}}$, we have

$$\mathfrak{R}^T \leq 6\sqrt{XAH^3T\iota} + HXA\iota.$$

Additionally, recall the naive bound $\mathfrak{R}^T \leq HT$ on the regret (which follows as $\langle \mu^t, \ell^t \rangle \in [0, H]$ for any $\mu \in \Pi_{\max}$, $t \in [T]$), we get

$$\mathfrak{R}^T \leq \min\left\{6\sqrt{XAH^3T\iota} + HXA\iota, HT\right\} \leq HT \cdot \min\left\{6\sqrt{XAH\iota/T} + XA\iota/T, 1\right\}.$$

For $T > HXA\iota$, the min above is upper bounded by $7\sqrt{HXA\iota/T}$. For $T \leq HXA\iota$, the min above is upper bounded by $1 \leq 7\sqrt{HXA\iota/T}$. Therefore, we always have

$$\mathfrak{R}^T \leq HT \cdot 7\sqrt{HXA\iota/T} = 7\sqrt{H^3XAT\iota}.$$

This is the desired result. □

The rest of this section is devoted to proving the above three lemmas.

E.3 A CONCENTRATION RESULT

We begin by presenting a useful concentration result. This result is a variant of (Kozuno et al., 2021, Lemma 3) and (Neu, 2015, Lemma 1) suitable to our loss estimator (8) where the IX bonus on the denominator depends on (x_h, a_h).

Lemma E.5. *For some fixed* $h \in [H]$*, let* $\alpha_h^t(x_h, a_h) \in \left[0, 2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)\right]$ *be* $\mathcal{F}^{t-1}$*-measurable random variable for each* $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$*. Then with probability* $1-\delta$*,*

$$\sum_{t=1}^{T} \sum_{x_h, a_h} \alpha_h^t(x_h, a_h)\left(\widetilde{\ell}_h^t(x_h, a_h) - \ell_h^t(x_h, a_h)\right) \leq \log(1/\delta).$$

Proof. Define the unbiased importance sampling estimator

$$\widehat{\ell}_h^t := \frac{1 - r_h^t}{\mu_{1:h}^t(x_h^t, a_h^t)} \cdot \mathbf{1}\left\{x_h = x_h^t, a_h = a_h^t\right\}.$$

We first have

$$\begin{aligned}
\widetilde{\ell}_h^t(x_h, a_h) =& \frac{1 - r_h^t}{\mu_{1:h}^t(x_h, a_h) + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)} \cdot \mathbf{1}\left\{x_h = x_h^t, a_h = a_h^t\right\} \\
\leq& \frac{1 - r_h^t}{\mu_{1:h}^t(x_h, a_h) + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)(1 - r_h^t)} \cdot \mathbf{1}\left\{x_h = x_h^t, a_h = a_h^t\right\} \\
\leq& \frac{1}{2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)} \frac{2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)(1 - r_h^t)\mathbf{1}\left\{x_h = x_h^t, a_h = a_h^t\right\}/\mu_{1:h}^t(x_h, a_h)}{1 + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)(1 - r_h^t)\mathbf{1}\left\{x_h = x_h^t, a_h = a_h^t\right\}/\mu_{1:h}^t(x_h, a_h)} \\
=& \frac{1}{2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)} \frac{2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)\widehat{\ell}_h^t(x_h, a_h)}{1 + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)\widehat{\ell}_h^t(x_h, a_h)} \\
\overset{(i)}{\leq}& \frac{1}{2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)} \log\left(1 + 2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)\widehat{\ell}_h^t(x_h, a_h)\right),
\end{aligned}$$

where (i) is because for any $z \geq 0$, $\frac{z}{1+z/2} \leq \log(1+z)$.

As a result, we have the following bound on the moment generating function:

$$
\begin{aligned}
&\mathbb{E}\left\{\exp\left\{\sum_{x_h,a_h} \alpha_h^t(x_h,a_h)\,\widetilde{\ell}_h^t(x_h,a_h)\right\}|\mathcal{F}^{t-1}\right\}\\
\leq&\mathbb{E}\left\{\exp\left\{\sum_{x_h,a_h} \frac{\alpha_h^t(x_h,a_h)}{2\gamma\mu_{1:h}^{\star,h}(x_h,a_h)}\log\left(1+2\gamma\mu_{1:h}^{\star,h}(x_h,a_h)\widehat{\ell}_h^t(x_h,a_h)\right)\right\}|\mathcal{F}^{t-1}\right\}\\
\overset{(i)}{\leq}&\mathbb{E}\left\{\exp\left\{\sum_{x_h,a_h} \log\left(1+\alpha_h^t(x_h,a_h)\,\widehat{\ell}_h^t(x_h,a_h)\right)\right\}|\mathcal{F}^{t-1}\right\}\\
=&\mathbb{E}\left\{\prod_{x_h,a_h}\left(1+\alpha_h^t(x_h,a_h)\,\widehat{\ell}_h^t(x_h,a_h)\right)|\mathcal{F}^{t-1}\right\}\\
\overset{(ii)}{=}&\mathbb{E}\left\{1+\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\,\widehat{\ell}_h^t(x_h,a_h)|\mathcal{F}^{t-1}\right\}\\
=&1+\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\,\ell_h^t(x_h,a_h)\\
\leq&\mathbb{E}\left\{\exp\left\{\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\,\ell_h^t(x_h,a_h)\right\}|\mathcal{F}^{t-1}\right\},
\end{aligned}
$$

where (i) is because $z\log(1+z') \leq \log(1+zz')$ for any $0 \leq z \leq 1$ and $z' > -1$, and (ii) follows from the fact that for any h, at most one of $\widehat{\ell}_h^t(x_h,a_h)$ is non-zero, so the cross terms disappear.

Repeating the above argument,

$$
\begin{aligned}
&\mathbb{E}\left\{\exp\left\{\sum_{t=1}^{T}\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\left(\widetilde{\ell}_h^t(x_h,a_h)-\ell_h^t(x_h,a_h)\right)\right\}\right\}\\
\leq&\mathbb{E}\left\{\exp\left\{\sum_{t=1}^{T-1}\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\left(\widetilde{\ell}_h^t(x_h,a_h)-\ell_h^t(x_h,a_h)\right)\right\}\mathbb{E}\left\{\exp\left\{\sum_{x_h,a_h}\alpha_h^T(x_h,a_h)\left(\widetilde{\ell}_h^T(x_h,a_h)-\ell_h^T(x_h,a_h)\right)\right\}|\mathcal{F}\right.\right.\\
\leq&\mathbb{E}\left\{\exp\left\{\sum_{t=1}^{T-1}\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\left(\widetilde{\ell}_h^t(x_h,a_h)-\ell_h^t(x_h,a_h)\right)\right\}\right\}\\
\leq&\cdots\leq 1.
\end{aligned}
$$

Therefore, we can apply the Markov inequality and get

$$
\begin{aligned}
&\mathbb{P}\left\{\sum_{t=1}^{T}\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\left(\widetilde{\ell}_h^t(x_h,a_h)-\ell_h^t(x_h,a_h)\right)>\log(1/\delta)\right\}\\
=&\mathbb{P}\left\{\exp\left\{\sum_{t=1}^{T-1}\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\left(\widetilde{\ell}_h^t(x_h,a_h)-\ell_h^t(x_h,a_h)\right)\right\}>1/\delta\right\}\\
\leq&\delta\cdot\mathbb{E}\left\{\exp\left\{\sum_{t=1}^{T}\sum_{x_h,a_h}\alpha_h^t(x_h,a_h)\left(\widetilde{\ell}_h^t(x_h,a_h)-\ell_h^t(x_h,a_h)\right)\right\}\right\}\leq\delta.
\end{aligned}
$$

This is the desired result. □

Corollary E.6. *We have*

(a) For some fixed $h \in [H]$ and (x_h, a_h), let $\alpha_h^t(x_h, a_h) \in \left[0, 2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)\right]$ be $\mathcal{F}^{t-1}$-measurable random variable. Then with probability $1-\delta$,

$$\sum_{t=1}^{T} \alpha_h^t(x_h, a_h)\left(\widetilde{\ell}_h^t(x_h, a_h) - \ell_h^t(x_h, a_h)\right) \le \log(1/\delta).$$

(b) For some fixed $h \in [H]$ and x_h, let $\alpha_h^t(x_h, a_h) \in \left[0, 2\gamma\mu_{1:h}^{\star,h}(x_h, a_h)\right]$ be $\mathcal{F}^{t-1}$-measurable random variable for each $a_h \in \mathcal{A}$. Then with probability $1-\delta$,

$$\sum_{t=1}^{T} \sum_{a_h \in \mathcal{A}} \alpha_h^t(x_h, a_h)\left(\widetilde{\ell}_h^t(x_h, a_h) - \ell_h^t(x_h, a_h)\right) \le \log(1/\delta).$$

Proof. For (a), using Lemma E.5 with $(\alpha_h^t)'(x_h', a_h') = \alpha_h^t(x_h', a_h')\,\mathbf{1}\{x_h' = x_h, a_h' = a_h\}$,

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_h^t(x_h, a_h)\left(\widetilde{\ell}_h^t(x_h, a_h) - \ell_h^t(x_h, a_h)\right) \\
=&\sum_{t=1}^{T} \sum_{x_h', a_h'} \alpha_h^t(x_h, a_h)\,\mathbf{1}\{x_h' = x_h, a_h' = a_h\}\left[\widetilde{\ell}^t(x_h', a_h') - \ell^t(x_h', a_h')\right] \le \log(1/\delta).
\end{aligned}$$

Claim (b) can proved similarly. □

E.4 Proof of Lemma E.2

We further decompose BIAS^1 to two terms by

$$\text{BIAS}^1 = \sum_{t=1}^{T}\left\langle \mu^t, \ell^t - \widetilde{\ell}^t\right\rangle = \underbrace{\sum_{t=1}^{T}\left\langle \mu^t, \ell^t - \mathbb{E}\left\{\widetilde{\ell}^t | \mathcal{F}^{t-1}\right\}\right\rangle}_{(A)} + \underbrace{\sum_{t=1}^{T}\left\langle \mu^t, \mathbb{E}\left\{\widetilde{\ell}^t | \mathcal{F}^{t-1}\right\} - \widetilde{\ell}^t\right\rangle}_{(B)}.$$

To bound (A), plug in the definition of loss estimator,

$$\begin{aligned}
&\sum_{t=1}^{T}\left\langle \mu^t, \ell^t - \mathbb{E}\left\{\widetilde{\ell}^t | \mathcal{F}^{t-1}\right\}\right\rangle \\
=&\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{x_h, a_h} \mu_{1:h}^t(x_h, a_h)\left[\ell_h^t(x_h, a_h) - \frac{\mu_{1:h}^t(x_h, a_h)\ell_h^t(x_h, a_h)}{\mu_{1:h}^t(x_h, a_h) + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)}\right] \\
=&\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{x_h, a_h} \mu_{1:h}^t(x_h, a_h)\ell_h^t(x_h, a_h)\left[\frac{\gamma\mu_{1:h}^{\star,h}(x_h, a_h)}{\mu_{1:h}^t(x_h, a_h) + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)}\right] \\
\le&\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{x_h, a_h} \gamma\mu_{1:h}^{\star,h}(x_h, a_h)\ell_h^t(x_h, a_h) \\
\overset{(i)}{\le}&\sum_{t=1}^{T}\sum_{h=1}^{H}\gamma = \gamma HT,
\end{aligned}$$

where (i) is by using Corollary D.2 with policy $\mu^{\star,h}$ for each layer h.

To bound (B), first notice

$$\left\langle \mu^t, \widetilde{\ell}^t\right\rangle = \sum_{h=1}^{H}\sum_{x_h, a_h} \mu_{1:h}^t(x_h, a_h)\frac{(1-r_h^t)\,\mathbf{1}\{x_h = x_h^t, a_h = a_h^t\}}{\mu_{1:h}^t(x_h, a_h) + \gamma\mu_{1:h}^{\star,h}(x_h, a_h)}$$

$$\leq \sum_{h=1}^{H} \sum_{x_h, a_h} \mathbf{1}\left\{x_h = x_h^t, a_h = a_h^t\right\} = \sum_{h=1}^{H} 1 = H.$$

Then by Azuma-Hoeffding, with probability at least $1 - \delta/3$,

$$\sum_{t=1}^{T} \left\langle \mu^t, \mathbb{E}\left\{\widetilde{\ell}^t | \mathcal{F}^{t-1}\right\} - \widetilde{\ell}^t \right\rangle \leq H\sqrt{2T\log(3/\delta)} \leq H\sqrt{2T\iota}.$$

Combining the bounds for (A) and (B) gives the desired result. □

E.5 Proof of Lemma E.3

We have

$$\begin{aligned}
\mathrm{BIAS}^2 &= \max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \left\langle \mu^\dagger, \widetilde{\ell}^t - \ell^t \right\rangle \\
&= \max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{x_h, a_h} \mu^\dagger_{1:h}(x_h, a_h) \left[\widetilde{\ell}^t_h(x_h, a_h) - \ell^t_h(x_h, a_h)\right] \\
&= \max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{x_h, a_h} \frac{\mu^\dagger_{1:h}(x_h, a_h)}{\gamma \mu^{\star,h}_{1:h}(x_h, a_h)} \gamma \mu^{\star,h}_{1:h}(x_h, a_h) \left[\widetilde{\ell}^t_h(x_h, a_h) - \ell^t_h(x_h, a_h)\right] \\
&= \max_{\mu^\dagger \in \Pi_{\max}} \sum_{h=1}^{H} \sum_{x_h, a_h} \frac{\mu^\dagger_{1:h}(x_h, a_h)}{\gamma \mu^{\star,h}_{1:h}(x_h, a_h)} \sum_{t=1}^{T} \gamma \mu^{\star,h}_{1:h}(x_h, a_h) \left[\widetilde{\ell}^t_h(x_h, a_h) - \ell^t_h(x_h, a_h)\right] \\
&\overset{(i)}{\leq} \frac{\log(XA/\delta)}{\gamma} \sum_{h=1}^{H} \max_{\mu^\dagger \in \Pi_{\max}} \sum_{x_h, a_h} \frac{\mu^\dagger_{1:h}(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \\
&\overset{(ii)}{\leq} \frac{\iota}{\gamma} \sum_{h=1}^{H} X_h A = XA\iota/\gamma,
\end{aligned}$$

where (i) is by applying Corollary E.6 for each (x_h, a_h) pair and taking union bound, and (ii) is by Lemma D.4. □

E.6 Proof of Lemma E.4

We begin by stating the following lemma, which roughly speaking relates the task of bounding the regret to bounding the term $\left\langle \mu, \widetilde{\ell}^t \right\rangle + \frac{1}{\eta \mu^{\star,1}_{1:1}(x^t_1, a_1)} \log Z^t_1$.

Lemma E.7. *For any policy $\mu \in \Pi_{\max}$,*

$$\mathrm{D}^{\mathrm{bal}}(\mu \| \mu^{t+1}) - \mathrm{D}^{\mathrm{bal}}(\mu \| \mu^t) = \eta \left\langle \mu, \widetilde{\ell}^t \right\rangle + \frac{1}{\mu^{\star,1}_{1:1}(x^t_1, a_1)} \log Z^t_1.$$

Proof. By definition of $\mathrm{D}^{\mathrm{bal}}$ and the conditional form update rule in Algorithm 1,

$$\begin{aligned}
&\mathrm{D}^{\mathrm{bal}}(\mu \| \mu^{t+1}) - \mathrm{D}^{\mathrm{bal}}(\mu \| \mu^t) \\
&= \sum_{h=1}^{H} \sum_{x_h, a_h} \frac{\mu_{1:h}(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \log \frac{\mu^t_h(a_h | x_h)}{\mu^{t+1}_h(a_h | x_h)} \\
&= \sum_{h=1}^{H} \sum_{a_h} \frac{\mu_{1:h}(x^t_h, a_h)}{\mu^{\star,h}_{1:h}(x^t_h, a_h)} \log \frac{\mu^t_h(a_h | x^t_h)}{\mu^{t+1}_h(a_h | x^t_h)} \\
&= \sum_{h=1}^{H} \frac{\mu_{1:h}(x^t_h, a^t_h)}{\mu^{\star,h}_{1:h}(x^t_h, a^t_h)} \left[\eta \mu^{\star,h}_{1:h}(x^t_h, a^t_h) \widetilde{\ell}^t_h - \frac{\mu^{\star,h}_{1:h}(x^t_h, a^t_h)}{\mu^{\star,h+1}_{1:h+1}(x^t_{h+1}, a^t_{h+1})} \log Z^t_{h+1}\right] + \sum_{h=1}^{H} \sum_{a_h} \frac{\mu_{1:h}(x^t_h, a_h)}{\mu^{\star,h}_{1:h}(x^t_h, a_h)} \log Z^t_h \\
&= \eta \sum_{h=1}^{H} \mu_{1:h}(x^t_h, a^t_h) \widetilde{\ell}^t_h(x^t_h, a^t_h) - \sum_{h=1}^{H} \frac{\mu_{1:h}(x^t_h, a^t_h)}{\mu^{\star,h+1}_{1:h+1}(x^t_{h+1}, a^t_{h+1})} \log Z^t_{h+1} + \sum_{h=1}^{H} \frac{\mu_{1:h-1}(x^t_{h-1}, a^t_{h-1})}{\mu^{\star,h}_{1:h}(x^t_h, a^t_h)} \log Z^t_h
\end{aligned}$$

$$=\eta\left\langle\mu,\widetilde{\ell}^t\right\rangle+\frac{1}{\mu_{1:1}^{\star,1}(x_1^t,a_1)}\log Z_1^t.$$

□

Additional notation We introduce the following notation for convenience throughout the rest of this subsection. Define

$$\beta_h^t := \eta\mu_{1:h}^{\star,h}(x_h^t,a_h^t).$$

For simplicity, when there is no confusion, we write

$$\mu_h^t := \mu_h^t(a_h^t|x_h^t),\quad \mu_{h:h'}^t := \prod_{h''=h}^{h'}\mu_{h''}^t,$$

and

$$\widetilde{\ell}_h^t := \widetilde{\ell}_h^t\left(x_h^t,a_h^t\right)=\frac{1-r_h^t}{\mu_{1:h}^t(x_h^t,a_h^t)+\gamma\mu_{1:h}^\star(x_h^t,a_h^t)}.$$

Define the normalized log-partition function as

$$\Xi_h^t := \frac{1}{\beta_h^t}\log Z_h^t=\frac{1}{\beta_h^t}\log\left(1-\mu_h^t+\mu_h^t\exp\left[\beta_h^t\left(\Xi_{h+1}^t-\widetilde{\ell}_h^t\right)\right]\right).$$

Note that this value can be seen as an H-variate function of the loss estimator $\left\{\widetilde{\ell}_h^t\right\}_{h\in[H]}$. To make this dependence more clear, for any $\widetilde{\ell}\in[0,\infty)^H$, we define the function $\{\Xi_h^t(\cdot)\}_{h=1}^H$ recursively by (overloading notation)

$$\Xi_h^t\left(\widetilde{\ell}\right)=\Xi_h^t\left(\widetilde{\ell}_{h:H}\right):=\begin{cases}\log\left(1-\mu_h^t+\mu_h^t\exp\left[-\beta_h^t\widetilde{\ell}_h\right]\right)/\beta_h^t & \text{if } h=H,\\ \log\left(1-\mu_h^t+\mu_h^t\exp\left[\beta_h^t\left(\Xi_{h+1}\left(\widetilde{\ell}_{h+1:H}\right)-\widetilde{\ell}_h\right)\right]\right)/\beta_h^t & \text{otherwise.}\end{cases}$$

With this definition, we have $\Xi_h^t=\Xi_h^t\left(\widetilde{\ell}^t\right)$ where $\widetilde{\ell}^t$ is the actual loss estimator. Note that, importantly, $\Xi_h^t(\widetilde{\ell}_{h:H})$ has a *compositional structure*: It is a function of $\widetilde{\ell}_h$ (h-th entry of the loss) and Ξ_{h+1}^t (which is itself a function of $\widetilde{\ell}_{h+1:H}$). This compositional structure is key to proving bounds on its gradients and Hessians.

The rest of this subsection is organized as follows. In Section E.6.1, we bound the gradients and Hessians of the function $\Xi_1^t(\cdot)$ in an entry-wise fashion, and then use the Mean-Value Theorem to give a bound on $\Xi_1^t=\Xi_1^t(\widetilde{\ell}^t)$ (Lemma E.11). We then combine this result with Lemma E.7 to prove the main lemma that bounds REGRET (Section E.6.2).

E.6.1 Bounding Ξ_1^t

Lemma E.8. *For $\widetilde{\ell}\in[0,\infty)^H$ and any $h\in[H]$, $\Xi_h^t\left(\widetilde{\ell}\right)\le 0$. Furthermore, $\Xi_h^t(0)=0$.*

Proof. We show the first claim by backward induction. For $h=H$,

$$\Xi_H^t\left(\widetilde{\ell}_H\right)=\log\left(1-\mu_H^t+\mu_H^t\exp\left[-\beta_H^t\widetilde{\ell}_H\right]\right)/\beta_H^t\le\log\left(1-\mu_H^t+\mu_H^t\right)/\beta_H^t\le 0,$$

because $\widetilde{\ell}_H^t\ge 0$.

Assume $\Xi_{h+1}^t\left(\widetilde{\ell}\right)\le 0$, then for the previous step h,

$$\Xi_h^t\left(\widetilde{\ell}_{h:H}\right)=\log\left(1-\mu_h^t+\mu_h^t\exp\left[\beta_h^t\left(\Xi_{h+1}^t\left(\widetilde{\ell}_{h+1:H}\right)-\widetilde{\ell}_h\right)\right]\right)/\beta_h^t\le\log\left(1-\mu_h^t+\mu_h^t\right)/\beta_h^t\le 0.$$

The second claim follows as all inequalities become equalities at $\widetilde{\ell}=0$. □

Lemma E.9 (Bounds on first derivatives)**.** *For $\widetilde{\ell} \in [0,1]^H$ and any $h \in [H]$, the derivatives are bounded by*

$$0 \leq \frac{\partial \Xi_h^t}{\partial \Xi_{h+1}^t} \leq \mu_h^t \quad \text{and} \quad -\mu_h^t \leq \frac{\partial \Xi_h^t}{\partial \widetilde{\ell}_h} \leq 0.$$

Furthermore,

$$\left.\frac{\partial \Xi_h^t}{\partial \widetilde{\ell}_{h'}}\right|_{\widetilde{\ell}=0} = \begin{cases} -\mu_{h:h'}^t & \text{if } h' \geq h, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. By chain rule and the compositional structure of the functions $\Xi_h^t(\cdot)$, for any $h' \geq h$,

$$\frac{\partial \Xi_h^t}{\partial \widetilde{\ell}_{h'}} = \frac{\partial \Xi_h^t}{\partial \Xi_{h'}^t} \cdot \frac{\partial \Xi_{h'}^t}{\partial \widetilde{\ell}_{h'}} = \left(\prod_{h''=h}^{h'-1} \frac{\partial \Xi_{h''}^t}{\partial \Xi_{h''+1}^t} \right) \cdot \frac{\partial \Xi_{h'}^t}{\partial \widetilde{\ell}_{h'}}.$$

For any h, the derivatives are bounded by

$$\frac{\partial \Xi_h^t}{\partial \Xi_{h+1}^t} = \frac{\mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]}{1 - \mu_h^t + \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]} \in \left[0, \mu_h^t\right],$$

$$\frac{\partial \Xi_h^t}{\partial \widetilde{\ell}_h} = -\frac{\mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]}{1 - \mu_h^t + \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]} \in \left[-\mu_h^t, 0\right].$$

The inequalities hold because the function $f(z) = \frac{\mu_h^t z}{1-\mu_h^t+\mu_h^t z} = 1 - \frac{1-\mu_h^t}{1-\mu_h^t+\mu_h^t z}$ is increasing on $z \in [0,1]$, and $\exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right] \in [0,1]$ by Lemma E.8.

Putting them together, at $\widetilde{\ell} = 0$, the derivative is just $\left.\frac{\partial \Xi_h^t}{\partial \widetilde{\ell}_{h'}}\right|_{\widetilde{\ell}^t=0} = -\mu_{h:h'}^t$ if $h' \geq h$. If $h' < h$, since Ξ_h^t only depends on loss in the later layers, $\frac{\partial \Xi_h^t}{\partial \widetilde{\ell}_{h'}}|_{\widetilde{\ell}^t=0} = 0$. □

Lemma E.10 (Bounds on second derivatives)**.** *For $\widetilde{\ell} \in [0,1]^H$ and any $h \in [H]$, if $h' \geq h$ and $h'' \geq h$, the second-order derivatives are bounded by*

$$\frac{\partial^2 \Xi_h^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}} \leq \sum_{h'''=h}^{\min\{h',h''\}} \beta_{h'''}^t \mu_{h:h'}^t \mu_{h'''+1:h''}^t = \sum_{h'''=h}^{\min\{h',h''\}} \beta_{h'''}^t \mu_{h:h'''}^t \mu_{h'''+1:h'}^t \mu_{h'''+1:h''}^t.$$

Otherwise $\frac{\partial^2 \Xi_h^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}} = 0$.

Proof. By symmetry of the second derivatives and the right-hand side with respect to h' and h'', it suffices to prove the claim for $h'' \geq h'$ only.

By chain rule and the compositional structure of the functions $\Xi_h^t(\cdot)$,

$$\frac{\partial^2 \Xi_h^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}} = \frac{\partial^2 \Xi_h^t}{\partial \Xi_{h'}^t \partial \widetilde{\ell}_{h''}} \cdot \frac{\partial \Xi_{h'}^t}{\partial \widetilde{\ell}_{h'}} + \frac{\partial \Xi_h^t}{\partial \Xi_{h'}^t} \cdot \frac{\partial^2 \Xi_{h'}^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}}.$$

If $h'' = h' = h$,

$$\frac{\partial^2 \Xi_h^t}{\partial \widetilde{\ell}_h^2} = \beta_h^t \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right] \frac{1 - \mu_h^t}{\left\{1 - \mu_h^t + \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]\right\}^2} \leq \beta_h^t \mu_h^t.$$

If $h' = h, h'' > h$,

$$\frac{\partial^2 \Xi_h^t}{\partial \widetilde{\ell}_h \partial \widetilde{\ell}_{h''}} = -\frac{(1-\mu_h^t)\beta_h^t \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]}{\left(1 - \mu_h^t + \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]\right)^2} \cdot \frac{\partial \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h''}} \leq \beta_h^t \mu_{h:h''}^t.$$

If $h < h' < h''$, we can compute the Hessian by induction. Notice once $h' > h$ we have

$$\frac{\partial \Xi_h^t}{\partial \widetilde{\ell}_{h'}} = \frac{\partial \Xi_h^t}{\partial \Xi_{h+1}^t} \cdot \frac{\partial \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h'}}.$$

Take second derivative,

$$\frac{\partial^2 \Xi_h^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}} = \underbrace{\frac{\partial \Xi_h^t}{\partial \Xi_{h+1}^t} \cdot \frac{\partial^2 \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}}}_{(i)} + \underbrace{\frac{\partial^2 \Xi_h^t}{\partial \Xi_{h+1}^t \partial \widetilde{\ell}_{h''}} \cdot \frac{\partial \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h'}}}_{(ii)}.$$

We first bound the second term,

$$\begin{aligned}
(ii) &= \frac{(1-\mu_h^t)\beta_h^t \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]}{\left(1 - \mu_h^t + \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]\right)^2} \cdot \frac{\partial \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h''}} \cdot \frac{\partial \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h'}} \\
&\leq \beta_h^t \mu_h^t \cdot \mu_{h+1:h''}^t \cdot \mu_{h+1:h'}^t \\
&\leq \beta_h^t \mu_{h:h'}^t \mu_{h+1:h''}^t.
\end{aligned}$$

The first term can be simplified to

$$(i) \leq \frac{\mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]}{1 - \mu_h^t + \mu_h^t \exp\left[\beta_h^t \left(\Xi_{h+1}^t\left(\widetilde{\ell}\right) - \widetilde{\ell}_h\right)\right]} \frac{\partial^2 \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}} \leq \mu_h^t \frac{\partial^2 \Xi_{h+1}^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}}.$$

Now plug in $\frac{\partial^2 \Xi_{h'}^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}} \leq \beta_{h'}^t \mu_{h':h''}^t$ and backward induction from h' to h gives:

$$\frac{\partial^2 \Xi_h^t}{\partial \widetilde{\ell}_{h'} \partial \widetilde{\ell}_{h''}} \leq \sum_{h'''=h}^{h'} \beta_{h'''}^t \mu_{h:h'}^t \mu_{h'''+1:h''}^t.$$

We can check this expression is also correct for the above special cases when $h' = h$. The second claim holds because Ξ_h^t only depends on loss in the later layers. □

Lemma E.11 (Bound on Ξ_1^t)**.** *We have*

$$\Xi_1^t \leq -\left\langle \mu^t, \widetilde{\ell}^t \right\rangle + \frac{\eta H}{2} \sum_{h=1}^{H} \left(\sum_{h'=h}^{H} \sum_{x_{h'}, a_{h'}} \mu_{1:h}^{\star,h}\left(x_{h'}, a_{h'}\right) \mu_{h+1:h'}^t\left(x_{h'}, a_{h'}\right) \widetilde{\ell}_{h'}^t\left(x_{h'}, a_{h'}\right) \right).$$

Proof. We apply the Mean-value Theorem to function $\Xi_1^t\left(\widetilde{\ell}\right)$ at $\widetilde{\ell} = 0$,

$$\Xi_1^t = \Xi_1^t\left(\widetilde{\ell}^t\right) = \Xi_1^t\left(0\right) + \left\langle \nabla_{\widetilde{\ell}} \Xi_1^t\big|_{\widetilde{\ell}=0}, \widetilde{\ell}^t \right\rangle + \frac{1}{2} \left\langle \nabla_{\widetilde{\ell}}^2 \Xi_1^t\big|_{\widetilde{\ell}=\xi^t} \widetilde{\ell}^t, \widetilde{\ell}^t \right\rangle,$$

where ξ^t lies on the line segment between 0 and $\widetilde{\ell}^t$.

By Lemma E.8, the initial term is just zero. By Lemma E.9, the first-order term is just $-\left\langle \mu^t, \widetilde{\ell}^t \right\rangle$.

It thus remains to bound the second-order term. Applying the entry-wise upper bounds in Lemma E.10 at $h = 1$ (which hold uniformly at all nonnegative loss values, including ξ^t), we have

$$\left\langle \nabla_{\widetilde{\ell}}^2 \Xi_1^t\big|_{\widetilde{\ell}=\xi^t} \widetilde{\ell}^t, \widetilde{\ell}^t \right\rangle = \sum_{h=1}^{H} \sum_{h'=1}^{H} \frac{\partial^2 \Xi_1^t}{\partial \widetilde{\ell}_h \partial \widetilde{\ell}_{h'}}\bigg|_{\widetilde{\ell}=\xi^t} \widetilde{\ell}_h^t \widetilde{\ell}_{h'}^t$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \sum_{h=1}^{H} \sum_{h'=1}^{H} \sum_{h''=1}^{\min\{h,h'\}} \beta_{h''}^{t} \mu_{1:h}^{t} \mu_{h''+1:h'}^{t} \widetilde{\ell}_{h}^{t} \widetilde{\ell}_{h'}^{t} \\
&= \sum_{h=1}^{H} \mu_{1:h}^{t} \widetilde{\ell}_{h}^{t} \sum_{h'=1}^{H} \sum_{h''=1}^{\min\{h,h'\}} \beta_{h''}^{t} \mu_{h''+1:h'}^{t} \widetilde{\ell}_{h'}^{t} \\
&\overset{(ii)}{\leq} H \max_{h\in[H]} \sum_{h'=1}^{H} \sum_{h''=1}^{\min\{h,h'\}} \beta_{h''}^{t} \mu_{h''+1:h'}^{t} \widetilde{\ell}_{h'}^{t} \\
&= H \sum_{h'=1}^{H} \sum_{h''=1}^{h'} \beta_{h''}^{t} \mu_{h''+1:h'}^{t} \widetilde{\ell}_{h'}^{t} \\
&= H \sum_{h''=1}^{H} \sum_{h'=h''}^{H} \beta_{h''}^{t} \mu_{h''+1:h'}^{t} \widetilde{\ell}_{h'}^{t} \\
&= \eta H \sum_{h''=1}^{H} \left(\sum_{h'=h''}^{H} \mu_{1:h''}^{\star,h''} \left(x_{h'}^{t}, a_{h'}^{t}\right) \mu_{h''+1:h'}^{t} \widetilde{\ell}_{h'}^{t} \right) \\
&\overset{(iii)}{=} \eta H \sum_{h''=1}^{H} \left(\sum_{h'=h''}^{H} \sum_{x_{h'},a_{h'}} \mu_{1:h''}^{\star,h''} \left(x_{h'}, a_{h'}\right) \mu_{h''+1:h'}^{t} \left(x_{h'}, a_{h'}\right) \widetilde{\ell}_{h'}^{t}(x_{h'}, a_{h'}) \right),
\end{aligned}
$$

where (i) is by Lemma E.10; (ii) follows from the bound

$$
\sum_{h=1}^{H} \mu_{1:h}^{t} \widetilde{\ell}_{h}^{t} = \sum_{h=1}^{H} \mu_{1:h}^{t} \cdot \frac{1 - r_{h}^{t}}{\mu_{1:h}^{t} + \gamma \mu_{1:h}^{\star,h}} \leq H;
$$

and (iii) is because $\widetilde{\ell}_{h'}^{t}(x_{h'}, a_{h'}) = 0$ at all $(x_{h'}, a_{h'}) \neq (x_{h'}^{t}, a_{h'}^{t})$. □

Lemma E.12. *With probability at least* $1 - \delta/3$,

$$
\sum_{t=1}^{T} \Xi_{1}^{t} \leq -\sum_{t=1}^{T} \left\langle \mu^{t}, \widetilde{\ell}^{t} \right\rangle + \eta H^{3} T + \frac{\eta X A H^{2} \iota}{\gamma},
$$

where $\iota := \log(H/\delta)$.

Proof. Using Lemma E.11 and take the summation with respect to $t \in [T]$ we have

$$
\sum_{t=1}^{T} \Xi_{1}^{t} \leq -\sum_{t=1}^{T} \left\langle \mu^{t}, \widetilde{\ell}^{t} \right\rangle + \frac{\eta H}{2} \sum_{h=1}^{H} \sum_{h'=h}^{H} \sum_{t=1}^{T} \underbrace{\sum_{x_{h'},a_{h'}} \mu_{1:h}^{\star,h} \left(x_{h'}, a_{h'}\right) \mu_{h+1:h'}^{t} \left(x_{h'}, a_{h'}\right) \widetilde{\ell}_{h'}^{t} \left(x_{h'}, a_{h'}\right)}_{:=\Delta_{h,h'}^{t}}. \tag{28}
$$

Observe that the random variables $\Delta_{h,h'}^{t}$ satisfy the following:

- $\Delta_{h,h'}^{t} \leq X_{h'} A/\gamma$ almost surely:

$$
\begin{aligned}
\Delta_{h,h'}^{t} &= \sum_{x_{h'},a_{h'}} \mu_{1:h}^{\star,h} \left(x_{h'}, a_{h'}\right) \mu_{h+1:h'}^{t} \left(x_{h'}, a_{h'}\right) \frac{\left(1 - r_{h'}^{t}\right) \mathbf{1}\left\{x_{h'} = x_{h'}^{t}, a_{h'} = a_{h'}^{t}\right\}}{\mu_{1:h'}^{t} \left(x_{h'}, a_{h'}\right) + \gamma \mu_{1:h'}^{\star,h'} \left(x_{h'}, a_{h'}\right)} \\
&\leq \frac{1}{\gamma} \sum_{x_{h'},a_{h'}} \frac{\mu_{1:h}^{\star,h} \left(x_{h'}, a_{h'}\right) \mu_{h+1:h'}^{t} \left(x_{h'}, a_{h'}\right)}{\mu_{1:h'}^{\star,h'} \left(x_{h'}, a_{h'}\right)} \overset{(i)}{\leq} \frac{X_{h'} A}{\gamma},
\end{aligned}
$$

  where (i) is by using Lemma D.4 with the mixture of $\mu^{\star,h}$ and μ^{t}.

- $\mathbb{E}[\Delta^t_{h,h'}|\mathcal{F}_{t-1}] \le 1$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information after iteration $t-1$:

$$\mathbb{E}[\Delta^t_{h,h'}|\mathcal{F}_{t-1}] = \sum_{x_{h'},a_{h'}} \mu^{\star,h}_{1:h}(x_{h'},a_{h'})\,\mu^t_{h+1:h'}(x_{h'},a_{h'})\,\ell^t_{h'}(x_{h'},a_{h'}) \overset{(i)}{\le} 1,$$

  where (i) is by using Corollary D.2 with the mixture policy of $\mu^{\star,h}$ and μ^t.

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

$$\begin{aligned}
\mathbb{E}[(\Delta^t_{h,h'})^2|\mathcal{F}_{t-1}] &\overset{(i)}{=} \sum_{x_{h'},a_{h'}} \left[\left(\mu^{\star,h}_{1:h}(x_{h'},a_{h'})\,\mu^t_{h+1:h'}(x_{h'},a_{h'})\,\frac{(1-r^t_{h'})\,\mathbf{1}\{x_{h'}=x^t_{h'},a_{h'}=a^t_{h'}\}}{\mu^t_{1:h'}(x_{h'},a_{h'})+\gamma\mu^{\star,h'}_{1:h'}(x_{h'},a_{h'})}\right)^2\right] \\
&\le \sum_{x_{h'},a_{h'}} \left(\frac{\mu^{\star,h}_{1:h}(x_{h'},a_{h'})\,\mu^t_{h+1:h'}(x_{h'},a_{h'})}{\mu^t_{1:h'}(x_{h'},a_{h'})+\gamma\mu^{\star,h'}_{1:h'}(x_{h'},a_{h'})}\right)^2 \mu^t_{1:h'}(x_{h'},a_{h'}) \\
&\le \frac{1}{\gamma}\sum_{x_{h'},a_{h'}} \frac{\mu^{\star,h}_{1:h}(x_{h'},a_{h'})\,\mu^t_{h+1:h'}(x_{h'},a_{h'})}{\mu^{\star,h'}_{1:h'}(x_{h'},a_{h'})} \overset{(ii)}{\le} \frac{X_{h'}A}{\gamma},
\end{aligned}$$

  where (i) follows from the fact that for any h, at most one of indicators is non-zero, so the cross terms disappear and (ii) is using Corollary D.2 with the mixture policy of $\mu^{\star,h}$ and μ^t.

Therefore, we can apply Freedman's inequality (Lemma B.1) and union bound to get that, with probability at least $1-\delta/3$, for some fixed $\lambda_{h,h'} \in (0, \gamma/X_{h'}A]$, the following holds simultaneously for all h, h':

$$\sum_{t=1}^T \Delta^t_{h,h'} \le \frac{\lambda_{h,h'}X_{h'}AT}{\gamma} + \frac{2\log(H/\delta)}{\lambda_{h,h'}} + T,$$

Take $\lambda_{h,h'} = \gamma/X_{h'}A$, we have

$$\sum_{t=1}^T \Delta^t_{h,h'} \le \frac{X_{h'}A\cdot 2\log(H/\delta)}{\gamma} + 2T.$$

Plug into equation (28), we have

$$\sum_{t=1}^T \Xi^t_1 \le -\sum_{t=1}^T \left\langle \mu^t, \widetilde{\ell}^t \right\rangle + \eta H^3 T + \frac{\eta H^2 X A\iota}{\gamma},$$

where $\iota := \log(H/\delta)$ is a log factor. □

E.6.2 Proof of main lemma

By Lemma E.7, for any policy $\mu^\dagger \in \Pi_{\max}$,

$$\frac{1}{\eta}\left(\mathrm{D}^{\mathrm{bal}}(\mu^\dagger\|\mu^{t+1}) - \mathrm{D}^{\mathrm{bal}}(\mu^\dagger\|\mu^t)\right) = \left\langle \mu^\dagger, \widetilde{\ell}^t \right\rangle + \Xi^t_1.$$

Taking the summation w.r.t. $t \in [T]$ and using Lemma E.12, we have with probability at least $1-\delta/3$, the following holds simultaneously over all $\mu^\dagger \in \Pi_{\max}$:

$$\begin{aligned}
\frac{1}{\eta}\left(\mathrm{D}^{\mathrm{bal}}(\mu^\dagger\|\mu^T) - \mathrm{D}^{\mathrm{bal}}(\mu^\dagger\|\mu^1)\right) &= \sum_{t=1}^T \left\langle \mu^\dagger, \widetilde{\ell}^t \right\rangle + \sum_{t=1}^T \Xi^t_1 \\
&\le \sum_{t=1}^T \left\langle \mu^\dagger - \mu^t, \widetilde{\ell}^t \right\rangle + \eta H^3 T + \frac{\eta H^2 X A\iota}{\gamma}.
\end{aligned}$$

Rerranging the terms we have

$$\max_{\mu^\dagger\in\Pi_{\max}} \sum_{t=1}^T \left\langle \mu^t - \mu^\dagger, \widetilde{\ell}^t \right\rangle \le \max_{\mu^\dagger\in\Pi_{\max}} \frac{1}{\eta}\left(\mathrm{D}^{\mathrm{bal}}(\mu^\dagger\|\mu^1) - \mathrm{D}^{\mathrm{bal}}(\mu^\dagger\|\mu^T)\right) + \eta H^3 T + \frac{\eta H^2 X A\iota}{\gamma}$$

$$\leq \max_{\mu^\dagger \in \Pi_{\max}} \frac{1}{\eta} \mathrm{D}^{\mathrm{bal}}(\mu^\dagger \| \mu^1) + \eta H^3 T + \frac{\eta H^2 X A \iota}{\gamma}$$
$$\leq \frac{X A \log A}{\eta} + \eta H^3 T + \frac{\eta H^2 X A \iota}{\gamma},$$

where the last inequality above follows by recalling that μ^1 is taken to be the uniform policy ($\mu^1_h(a_h|x_h) = 1/A$ for all (h, x_h, a_h)) in Algorithm 1, and applying the bound on the balanced dilated KL (Lemma D.7). This proves Lemma E.4.

F LOWER BOUND FOR LEARNING IIEFGS

Theorem F.1 (Lower bound for learning IIEFGs)**.** *For any $A \geq 2$, $H \geq 1$, we have ($c > 0$ is an absolute constant)*

(a) (Regret lower bound) For any algorithm that controls the max player and plays policies $\{\mu^t\}_{t=1}^T$ where $T \geq XA$, there exists a game with $B = 1$ on which

$$\mathbb{E}\left[\mathfrak{R}^T\right] = \mathbb{E}\left[\max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^T \left\langle \mu^t - \mu^\dagger, \ell^t \right\rangle\right] \geq c \cdot \sqrt{XAT}.$$

(b) (PAC lower bound for learning NE) For any algorithm that controls both players and outputs a final policy $(\widehat{\mu}, \widehat{\nu})$ with T episodes of play, and any $\varepsilon \in (0, 1]$, there exists a game on which the algorithm suffers from $\mathbb{E}[\mathrm{NEGap}(\widehat{\mu}, \widehat{\nu})] \geq \varepsilon$, unless

$$T \geq c \cdot (XA + YB)/\varepsilon^2.$$

The proof of Theorem F.1 constructs a hard instance with $X = \Theta(X_H) = \Theta(A^{H-1})$ that is equivalent to A^H-armed bandit problems, and follows by a reduction to standard bandit lower bounds. We remark that our lower bounds are tight in X but did not explicitly optimize the H dependence (which is typically lower-order compared to X).

Proof. Both the regret and PAC lower bounds follow from a direct reduction to stochastic multi-armed bandits. For completeness, we first state the lower bound for stochastic bandits (Lattimore & Szepesvári, 2020, Exercise 15.4 & Exercise 33.1) as follows. Below, c is an absolute constant.

Proposition F.2 (Lower bound for stochastic bandits)**.** *Let $K \geq 2$ denote the number of arms.*

(a) (Regret lower bound) Suppose $T \geq K$. For any bandit algorithm that plays policy $\mu^t \in \Delta([K])$ (either deterministic or random) in round $t \in [T]$, there exists some K-armed stochastic bandit problem with Bernoulli rewards with mean vector $r \in [0, 1]^K$, on which the algorithm suffers from the following lower bound on the expected regret:

$$\mathbb{E}\left[\max_{\mu^\dagger \in \Delta([K])} \sum_{t=1}^T \left\langle \mu^\dagger - \mu^t, r \right\rangle\right] \geq c \cdot \sqrt{KT}.$$

(b) (PAC lower bound) For any bandit algorithm that plays for t rounds and outputs some policy $\widehat{\mu} \in \Delta([K])$, there exists some K-armed stochastic bandit problem with Bernoulli rewards with some mean vector $r \in [0, 1]^K$, on which policy $\widehat{\mu}$ is at least ε away from optimal:

$$\mathbb{E}\left[\max_{\mu^\dagger \in \Delta([K])} \left\langle \mu^\dagger - \widehat{\mu}, r \right\rangle\right] \geq \varepsilon,$$

unless $T \geq cK/\varepsilon^2$.

We now construct a class of IIEFGs with $X_H = A^{H-1}$ (the minimal possible number of infosets), and show that any algorithm that solves this class of games will imply an algorithm for stochastic bandits with A^H arms with the same regret/PAC bounds, from which Theorem F.1 follows.

Our construction is as follows: For any $A \geq 2$ and $H \geq 1$, we let $S_h = A^{h-1}$ for all $h \in [H]$ (in particular, $S_1 = 1$) and $B = 1$ (so that there is no opponent effectively). By the tree structure,

each state is thus uniquely determined by all past actions $s_h = (a_1, \dots, a_{h-1})$, and the transition is deterministic: $((a_1, \dots, a_{h-1}), a_h) \in \mathcal{S}_h \times \mathcal{A}$ transits to $(a_1, \dots, a_h) \in \mathcal{S}_{h+1}$ with probability one. Further, we let $x_h = x(s_h) = s_h$, so that there is no partial observability, and thus $\mathcal{X}_h = \mathcal{S}_h$ for all h. Only the H-th layer yields a Bernoulli reward with some mean $r_{a_{1:H}} := \mathbb{E}[r_H(a_{1:H-1}, a_H)] \in [0, 1]$, for all $a_{1:H} \in \mathcal{X}_H$. The reward is zero within all previous layers.

Under this model, the expected reward under any policy $\mu \in \Pi_{\max}$ can be succinctly written as

$$\langle \mu, r \rangle = \sum_{(x_H, a_H) \in \mathcal{X}_H \times \mathcal{A}} \mu_{1:H}(x_H, a_H) \mathbb{E}[r_H(x_h, a_H)] = \sum_{a_{1:H} \in \mathcal{A}^H} \mu_{1:H}(a_{1:H}) r_{a_{1:H}}.$$

This expression coincides with the expression for the expected reward of an A^H-armed stochastic bandit problem.

Now, for any algorithm Alg achieving regret $\mathfrak{R}^T$ on IIEFGs, we claim we can use it to design an algorithm for solving any A^H-armed stochastic bandit problem with Bernoulli rewards, and achieve the same regret. Indeed, given any A^H-armed bandit problem, we rename its arms as a sequence $a_{1:H} = (a_1, \dots, a_H) \in \mathcal{A}^H$. Now, we instantiate an instance of Alg on a simulated IIEFG with the above structure. Whenever Alg plays policy $\mu^t \in \Pi_{\max}$, we query an arm $a_{1:H}$ using policy $\mu^t_{1:H}(\cdot) \in \Delta(\mathcal{A}^H)$ in the bandit problem. Then, upon receiving the reward r^t from the bandit problem, we give the feedback that the game transitted to infoset $a_{1:H}$ and yielded reward r^t. By the above equivalence, the regret $\mathfrak{R}^T$ within this simulated game is exactly the same as the regret for the bandit problem.

Therefore, for $T \geq A^H$, we can apply Proposition F.2(a) to show that for any such Alg, there exists one such IIEFG, on which

$$\mathbb{E}\left[\mathfrak{R}^T\right] \geq c \cdot \sqrt{A^H T} = c\sqrt{X_H A T} \geq c\sqrt{XAT},$$

where the last inequality follows from the fact that $X \leq X_H(1 + 1/A + 1/A^2 + \cdots) \leq X_H/(1 - 1/A) \leq 2X_H$ by perfect recall. This shows part (a).

Part (b) (PAC lower bound) follows similarly from Proposition F.2(b). Using the same reduction, we can show for any algorithm that controls both players and outputs policy $(\widehat{\mu}, \widehat{\nu}) \in \Pi_{\max} \times \Pi_{\min}$, there exists one such game of the above form (where only the max player affects the game) where the algorithm suffers from the PAC lower bound

$$\mathbb{E}[\mathrm{NEGap}(\widehat{\mu}, \widehat{\nu})] = \mathbb{E}\left[\max_{\mu \in \Pi_{\max}} V^{\mu^\dagger, \widehat{\nu}} - V^{\widehat{\mu}, \widehat{\nu}}\right] \geq \varepsilon$$

unless $T \geq cXA/\varepsilon^2$. The symmetrical construction for the min player implies that there exists some game on which $\mathbb{E}[\mathrm{NEGap}(\widehat{\mu}, \widehat{\nu})] \geq \varepsilon$ unless $T \geq cYB/\varepsilon^2$.

Therefore, if $T < c(XA + YB)/(2\varepsilon^2)$, at least one of $T \geq cXA/\varepsilon^2$ and $T \geq cYB/\varepsilon^2$ has to be false, for which we obtain a game where the expected duality gap is at least ε. This shows part (b). □

G Proofs for Section 4

G.1 Counterfactual regret decomposition

Define the immediate counterfactual regret at any $x_h \in \mathcal{X}_h$, $h \in [H]$ as

$$\mathfrak{R}^{\mathrm{imm},T}_h(x_h) = \max_{\mu^\dagger_h(\cdot|x_h)} \sum_{t=1}^T \left\langle \mu^t_h(\cdot|x_h) - \mu^\dagger_h(\cdot|x_h), L^t_h(x_h, \cdot) \right\rangle, \tag{29}$$

where $L^t_h(\cdot, \cdot)$ is the counterfactual loss function defined in (10):

$$L^t_h(x_h, a_h) := \ell^t_h(x_h, a_h) + \sum_{h'=h+1}^H \sum_{(x_{h'}, a_{h'}) \in \mathcal{C}_{h'}(x_h, a_h) \times \mathcal{A}} \mu^t_{(h+1):h'}(x_{h'}, a_{h'}) \ell^t_{h'}(x_{h'}, a_{h'}).$$

Lemma G.1 (Counterfactual regret decomposition)**.** *We have* $\widetilde{\mathfrak{R}}^T \le \sum_{h=1}^H \mathfrak{R}_h^T$*, where*

$$\begin{aligned}\mathfrak{R}_h^T &:= \sum_{x_1\in\mathcal{X}_1} \max_{a_1\in\mathcal{A}} \cdots \sum_{x_{h-1}\in\mathcal{C}(x_{h-2},a_{h-2})} \max_{a_{h-1}\in\mathcal{A}} \sum_{x_h\in\mathcal{C}(x_{h-1},a_{h-1})} \mathfrak{R}_h^{\mathrm{imm},T}(x_h),\\ &= \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1})\cdot \mathfrak{R}_h^{\mathrm{imm},T}(x_h).\end{aligned}$$

Proof. The bound $\widetilde{\mathfrak{R}}^T \le \sum_{h=1}^H \mathfrak{R}_h^T$ with the sum-max form expression for $\widetilde{\mathfrak{R}}_h^T$ has already implicitly appeared in the proof of (Zinkevich et al., 2007, Theorem 3), albeit with their slightly different formulation of extensive-form games (turn-based games with reward only in the last round). For completeness, here we provide a proof under our formulation.

We first show the bound with the μ form expression for $\mathfrak{R}_h^T$, which basically follows by a performance decomposition argument. We have

$$\begin{aligned}\widetilde{\mathfrak{R}}^T &= \max_{\mu^\dagger\in\Pi_{\max}} \left\langle \mu^t - \mu^\dagger, \ell^t\right\rangle\\ &= \max_{\mu^\dagger\in\Pi_{\max}} \sum_{h=1}^H \left\langle \mu_{1:h-1}^\dagger \mu_{h:H}^t - \mu_{1:h}^\dagger \mu_{h+1:H}^t, \ell^t\right\rangle\\ &\le \sum_{h=1}^H \underbrace{\max_{\mu^\dagger\in\Pi_{\max}} \left\langle \mu_{1:h-1}^\dagger \mu_{h:H}^t - \mu_{1:h}^\dagger \mu_{h+1:H}^t, \ell^t\right\rangle}_{:=\mathfrak{R}_h^T}.\end{aligned}$$

Note that each term $\mathfrak{R}_h^T$ measures the performance difference between $\mu_{1:h-1}^\dagger \mu_{h:H}^t$ and $\mu_{1:h}^\dagger \mu_{h+1:H}^t$:

$$\begin{aligned}\mathfrak{R}_h^T &= \max_{\mu^\dagger\in\Pi_{\max}} \mathbb{E}_{s_h\sim\mu_{1:h-1}^\dagger\times\nu^t}\left[\mathbb{E}_{a_h\sim\mu^t(\cdot|x_h)}\left[\sum_{h'=1}^H r_{h'}\right] - \mathbb{E}_{a_h\sim\mu^\dagger(\cdot|x_h)}\left[\sum_{h'=1}^H r_{h'}\right]\right]\\ &\overset{(i)}{=} \max_{\mu^\dagger\in\Pi_{\max}} \mathbb{E}_{s_h\sim\mu_{1:h-1}^\dagger\times\nu^t}\left[\mathbb{E}_{a_h\sim\mu^t(\cdot|x_h)}\left[\sum_{h'=h}^H r_{h'}\right] - \mathbb{E}_{a_h\sim\mu^\dagger(\cdot|x_h)}\left[\sum_{h'=h}^H r_{h'}\right]\right]\\ &\overset{(ii)}{=} \max_{\mu^\dagger\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:h-1}^\dagger(x_{h-1},a_{h-1})\cdot\left\langle \mu_h^t(\cdot|x_h) - \mu_h^\dagger(\cdot|x_h), L_h^t(x_h,\cdot)\right\rangle\\ &= \max_{\mu^\dagger\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:h-1}^\dagger(x_{h-1},a_{h-1})\cdot \mathfrak{R}_h^{\mathrm{imm},T}(x_h).\end{aligned}$$

Above, (i) follows as the rewards for the first $h-1$ steps are the same for the two expectations; (ii) follows by definition of the counterfactual loss function (cumulative loss multiplied by the opponent and environment's policy / transition probabilities, as well as the max player's own policy from step h onward). The claim (with the μ form expression) thus follows by renaming the dummy variable $\mu^\dagger$ as μ.

To verify that the second expression is equivalent to the first expression, it suffices to notice that the max over $\mu_{1:h-1}\in\Pi_{\max}$ consists of separable optimization problems over $\mu_{h'}(\cdot|x_{h'})$ over all $x_{h'}\in\mathcal{X}_{h'}$, $h'\le h-1$, due to the perfect recall assumption (different $(x_{h'},a_{h'})$ leads to disjoint subtrees). Therefore, we can rewrite the above as

$$\begin{aligned}\mathfrak{R}_h^T = &\sum_{x_1\in\mathcal{X}_1} \max_{\mu_1(\cdot|x_1)\in\Delta(\mathcal{A})} \sum_{a_1\in\mathcal{A}} \mu_1(a_1|x_1) \sum_{x_2\in\mathcal{C}(x_1,a_1)} \cdots\\ &\sum_{x_{h-1}\in\mathcal{C}(x_{h-2},a_{h-2})} \max_{\mu_{h-1}(\cdot|x_{h-1})\in\Delta(\mathcal{A})} \sum_{a_{h-1}\in\mathcal{A}} \mu_{h-1}(a_{h-1}|x_{h-1}) \sum_{x_h\in\mathcal{C}(x_{h-1},a_{h-1})} \mathfrak{R}_h^{\mathrm{imm},T}(x_h).\end{aligned}$$

Further noticing (backward recursively) that each max over the action distribution is achieved at a single action yields the claimed sum-max form expression. □

G.2 Proof of Theorem 6

Overview of proof Our proof follows the usual analysis pipeline for MCCFR algorithms that decomposes the overall regret $\widetilde{\mathfrak{R}}_T$ into combinations of immediate counterfactual regrets $\mathfrak{R}_h^{\mathrm{imm},T}(x_h)$, and bounds each by regret bounds (of the regret minimizer R_{x_h}) plus concentration terms. We adopt a sharp application of this pipeline by using a tight counterfactual regret decomposition (Lemma G.1), as well as using the balancing property of $\mu^{\star,h}$ which yields sharp bounds on both the regret and concentration terms (Lemma G.2-G.4).

We now prove our main theorem on the regret of the CFR algorithm.

By Lemma G.1, we have $\widetilde{\mathfrak{R}}^T \leq \sum_{h=1}^H \mathfrak{R}_h^T$, where for any $h \in [H]$ we have

$$
\begin{aligned}
\mathfrak{R}_h^T &= \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1})\mathfrak{R}_h^{\mathrm{imm},T}(x_h) \\
&= \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \max_{\mu_h^\dagger(\cdot|x_h)} \sum_{t=1}^T \left\langle \mu_h^t(\cdot|x_h) - \mu_h^\dagger(\cdot|x_h), L_h^t(x_h,\cdot)\right\rangle \\
&\leq \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \underbrace{\max_{\mu_h^\dagger(\cdot|x_h)} \sum_{t=1}^T \left\langle \mu_h^t(\cdot|x_h) - \mu_h^\dagger(\cdot|x_h), \widetilde{L}_h^t(x_h,\cdot)\right\rangle}_{:=\widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h)} \\
&\quad + \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \sum_{t=1}^T \left\langle \mu_h^t(\cdot|x_h), L_h^t(x_h,\cdot) - \widetilde{L}_h^t(x_h,\cdot)\right\rangle \\
&\quad + \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \max_{\mu_h^\dagger(\cdot|x_h)} \sum_{t=1}^T \left\langle \mu_h^\dagger(\cdot|x_h), \widetilde{L}_h^t(x_h,\cdot) - L_h^t(x_h,\cdot)\right\rangle \\
&\overset{(i)}{=} \underbrace{\max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1})\widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h)}_{:=\mathrm{REGRET}_h} \\
&\quad + \underbrace{\max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \sum_{t=1}^T \mu_h^t(a_h|x_h)\Big[L_h^t(x_h,a_h) - \widetilde{L}_h^t(x_h,a_h)\Big]}_{:=\mathrm{BIAS}_h^1} \\
&\quad + \underbrace{\max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h,a_h) \sum_{t=1}^T \Big[\widetilde{L}_h^t(x_h,a_h) - L_h^t(x_h,a_h)\Big]}_{:=\mathrm{BIAS}_h^2} \\
&= \mathrm{REGRET}_h + \mathrm{BIAS}_h^1 + \mathrm{BIAS}_h^2.
\end{aligned}
$$

Above, the simplification of the BIAS_h^2 part in (i) uses the fact that the inner max over $\mu_h^\dagger(\cdot|x_h)$ and the outer max over $\mu_{1:(h-1)}$ are separable and thus can be merged into a single max over $\mu_{1:h}$.

We now state three lemmas that bound each term above. Their proofs are deferred to Sections G.3-G.5.

Lemma G.2 (Bound on BIAS_h^1)**.** *For any sequence of opponents' policies $\nu^t \in \mathcal{F}_{t-1}$, using the estimator $\widetilde{L}_h$ in (12), with probability $1-\delta/10$, we have*

$$
\sum_{h=1}^H \mathrm{BIAS}_h^1 \leq 2\sqrt{H^3XAT\iota} + HX\iota,
$$

where $\iota = \log(10X/\delta)$.

Lemma G.3 (Bound on BIAS_h^2)**.** *For any sequence of opponents' policies $\nu^t \in \mathcal{F}_{t-1}$, using the estimator $\widetilde{L}_h$ in (12), with probability $1-\delta/10$, we have*

$$\sum_{h=1}^{H} \mathrm{BIAS}_h^2 \le 2\sqrt{H^3XAT\iota} + HXA\iota,$$

where $\iota = \log(10XA/\delta)$.

Lemma G.4 (Bound on REGRET_h)**.** *Choosing $\eta = \sqrt{XA\iota/(H^3T)}$, we have that with probability at least $1-\delta/10$ (over the randomness within the loss estimator $\widetilde{L}_h^t$),*

$$\sum_{h=1}^{H} \mathrm{REGRET}_h \le 2\sqrt{H^3XAT\iota} + \sqrt{HX^3A^3\iota^3/(4T)},$$

where $\iota = \log(10XA/\delta)$.

Combining Lemma G.2, G.3, and G.4, we obtain the following: Choosing $\eta = \sqrt{XA\iota/(H^3T)}$, with probability at least $1-3\delta/10 \ge 1-\delta$, we have

$$\begin{aligned}\widetilde{\mathfrak{R}}^T &\le \sum_{h=1}^{H} \mathfrak{R}_h^T \le \sum_{h=1}^{H} \mathrm{REGRET}_h + \sum_{h=1}^{H} \mathrm{BIAS}_h^1 + \sum_{h=1}^{H} \mathrm{BIAS}_h^2 \\ &\le 6\sqrt{H^3XAT\iota} + 2HXA\iota + \sqrt{HX^3A^3\iota^3/(4T)}.\end{aligned}$$

Additionally, recall the naive bound $\widetilde{\mathfrak{R}}^T \le HT$ on the regret (which follows as $\langle \mu^t, \ell^t \rangle \in [0, H]$ for any $\mu \in \Pi_{\max}$, $t \in [T]$), we get

$$\begin{aligned}\widetilde{\mathfrak{R}}^T &\le \min\left\{6\sqrt{H^3XAT\iota} + 2HXA\iota + \sqrt{HX^3A^3\iota^3/4T}, HT\right\} \\ &\le HT \cdot \min\left\{6\sqrt{HXA\iota/T} + 2XA\iota/T + \sqrt{X^3A^3\iota^3/(4HT^3)}, 1\right\}.\end{aligned}$$

For $T > HXA\iota$, the min above is upper bounded by $9\sqrt{HXA\iota/T}$. For $T \le HXA\iota$, the min above is upper bounded by $1 \le 9\sqrt{HXA\iota/T}$. Therefore, we always have

$$\widetilde{\mathfrak{R}}^T \le HT \cdot 9\sqrt{HXA\iota/T} = 9\sqrt{H^3XAT\iota}.$$

This is the desired result. □

G.3 Proof of Lemma G.2

Rewrite BIAS_h^1 as

$$\begin{aligned}\mathrm{BIAS}_h^1 &= \max_{\mu \in \Pi_{\max}} \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{\mu_{1:(h-1)}(x_{h-1}, a_{h-1})}{\mu^{\star,h}_{1:(h-1)}(x_{h-1}, a_{h-1})} \\ &\qquad \cdot \sum_{t=1}^{T} \mu^{\star,h}_{1:(h-1)}(x_{h-1}, a_{h-1}) \mu_h^t(a_h|x_h) \cdot \left[L_h^t(x_h, a_h) - \widetilde{L}_h^t(x_h, a_h)\right] \\ &= \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1}, a_{h-1})}{\mu^{\star,h}_{1:(h-1)}(x_{h-1}, a_{h-1})} \\ &\cdot \sum_{t=1}^{T} \underbrace{\sum_{a_h \in \mathcal{A}} \frac{\mu_h^t(a_h|x_h)}{\mu_h^{\star,h}(a_h|x_h)} \left[\mu^{\star,h}_{1:h}(x_h, a_h) L_h^t(x_h, a_h) - \left(H - h + 1 - \sum_{h'=h}^{H} r_{h'}^{t,(h)}\right) \mathbf{1}\left\{(x_h^{t,(h)}, a_h^{t,(h)}) = (x_h, a_h)\right\}\right]}_{:= \widetilde{\Delta}_t^{x_h}}.\end{aligned} \tag{30}$$

Observe that the random variables $\widetilde{\Delta}_t^{x_h}$ satisfy the following:

- $\widetilde{\Delta}_t^{x_h} \le H$ almost surely:

$$\begin{aligned}\widetilde{\Delta}_t^{x_h} &\le \sum_{a_h\in\mathcal{A}} \frac{\mu_h^t(a_h|x_h)}{\mu_h^{\star,h}(a_h|x_h)} \cdot \mu_{1:h}^{\star,h}(x_h,a_h) L_h^t(x_h,a_h) \\ &= \sum_{a_h\in\mathcal{A}} \mu_h^t(a_h|x_h)\mu_{1:(h-1)}^{\star,h}(x_{h-1},a_{h-1}) L_h^t(x_h,a_h) \le H.\end{aligned}$$

  Above, the last bound follows from Lemma D.1(a).

- $\mathbb{E}[\widetilde{\Delta}_t^{x_h}|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information after iteration $t-1$;

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

$$\begin{aligned}&\mathbb{E}\left[\left(\widetilde{\Delta}_t^{x_h}\right)^2\Big|\mathcal{F}_{t-1}\right] \\ &\le \mathbb{E}\left[\sum_{a_h\in\mathcal{A}}\left(\frac{\mu_h^t(a_h|x_h)}{\mu_h^{\star,h}(a_h|x_h)}\right)^2 \cdot \left(H-h+1-\sum_{h'=h}^{H} r_{h'}^{t,(h)}\right)^2 \mathbf{1}\left\{(x_h^{t,(h)},a_h^{t,(h)}) = (x_h,a_h)\right\}\Big|\mathcal{F}_{t-1}\right] \\ &\le H^2 \sum_{a_h\in\mathcal{A}}\left(\frac{\mu_h^t(a_h|x_h)}{\mu_h^{\star,h}(a_h|x_h)}\right)^2 \cdot \mathbb{P}^{\mu_{1:h}^{\star,h},\nu^t}\left((x_h^{t,(h)},a_h^{t,(h)}) = (x_h,a_h)\right) \\ &= H^2 \sum_{a_h\in\mathcal{A}}\left(\frac{\mu_h^t(a_h|x_h)}{\mu_h^{\star,h}(a_h|x_h)}\right)^2 \cdot \mu_{1:h}^{\star,h}(x_h,a_h)\cdot p_{1:h}^{\nu^t}(x_h) \\ &= H^2 \sum_{a_h\in\mathcal{A}}\underbrace{\left(\frac{\mu_h^t(a_h|x_h)}{\mu_h^{\star,h}(a_h|x_h)}\right)}_{\le A} \cdot \mu_{1:h-1}^{\star,h}(x_{h-1},a_{h-1})\cdot \mu_h^t(a_h|x_h) p_{1:h}^{\nu^t}(x_h) \\ &\le H^2 A \cdot \sum_{a_h\in\mathcal{A}} \mu_{1:h-1}^{\star,h}(x_{h-1},a_{h-1})\cdot \mu_h^t(a_h|x_h) p_{1:h}^{\nu^t}(x_h).\end{aligned}$$

Therefore, we can apply Freedman's inequality (Lemma B.1) and union bound to get that, for any fixed $\lambda \in (0, 1/H]$, with probability at least $1-\delta/10$, the following holds simultaneously for all (h, x_h):

$$\sum_{t=1}^T \widetilde{\Delta}_t^{x_h} \le \lambda H^2 A \sum_{a_h\in\mathcal{A}} \mu_{1:h-1}^{\star,h}(x_{h-1},a_{h-1})\cdot \sum_{t=1}^T \mu_h^t(a_h|x_h) p_{1:h}^{\nu^t}(x_h) + \frac{\iota}{\lambda},$$

where $\iota := \log(10X/\delta)$ is a log factor. Plugging this bound into (30) yields that, for all $h \in [H]$,

$$\begin{aligned}&\mathrm{BIAS}_h^1 = \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})}{\mu_{1:(h-1)}^{\star,h}(x_{h-1},a_{h-1})}\cdot \sum_{t=1}^T \widetilde{\Delta}_t^{x_h} \\ &\le \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})}{\mu_{1:(h-1)}^{\star,h}(x_{h-1},a_{h-1})}\cdot \left[\lambda H^2 A \sum_{a_h\in\mathcal{A}} \mu_{1:h-1}^{\star,h}(x_{h-1},a_{h-1})\cdot \sum_{t=1}^T \mu_h^t(a_h|x_h) p_{1:h}^{\nu^t}(x_h) + \frac{\iota}{\lambda}\right] \\ &\le \lambda H^2 A \cdot \max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h-1}(x_{h-1},a_{h-1}) \sum_{t=1}^T \mu_h^t(a_h|x_h) p_{1:h}^{\nu^t}(x_h) \\ &\quad + \frac{\iota}{\lambda}\cdot \max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})}{\mu_{1:(h-1)}^{\star,h}(x_{h-1},a_{h-1})} \\ &\overset{(i)}{=} \lambda H^2 AT + \frac{\iota}{\lambda}\cdot\frac{1}{A}\max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{(\mu_{1:(h-1)}\mu_h^{\mathrm{unif}})(x_h,a_h)}{\mu_{1:h}^{\star,h}(x_h,a_h)}\end{aligned}$$

$\overset{(ii)}{=} \lambda H^2 AT + \frac{\iota}{\lambda} \cdot X_h.$

Above, (i) used the fact that $\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h-1}(x_{h-1},a_{h-1})\mu_h^t(a_h|x_h)p_{1:h}^{\nu^t}(x_h) = 1$ for any $\mu \in \Pi_{\max}$ and any $t \in [T]$ (Lemma D.1(a)), as well as the fact that $\mu_h^{\star,h}(a_h|x_h) = \mu_h^{\text{unif}}(a_h|x_h) := 1/A$; (ii) used the balancing property of $\mu_{1:h}^{\star,h}$ (Lemma D.4). Combining the bounds for all $h \in [H]$, we get that with probability at least $1 - \delta/10$,

$$\sum_{h=1}^{H} \text{BIAS}_h^1 \le \lambda H^3 AT + \frac{X\iota}{\lambda}.$$

Choosing

$$\lambda = \min\left\{\sqrt{\frac{X\iota}{H^3 AT}}, \frac{1}{H}\right\} \le \frac{1}{H},$$

we obtain the bound

$$\sum_{h=1}^{H} \text{BIAS}_h^1 \le 2\sqrt{H^3 XAT\iota} + HX\iota.$$

This is the desired result. □

G.4 Proof of Lemma G.3

The proof strategy is similar to Lemma G.2. We can rewrite BIAS_h^2 as

$$\begin{aligned}
\text{BIAS}_h^2 &= \max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}(x_h,a_h)}{\mu_{1:h}^{\star,h}(x_h,a_h)} \cdot \sum_{t=1}^{T} \mu_{1:h}^{\star,h}(x_h,a_h)\left[\widetilde{L}_h^t(x_h,a_h) - L_h^t(x_h,a_h)\right] \\
&= \max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \frac{\mu_{1:h}(x_h,a_h)}{\mu_{1:h}^{\star,h}(x_h,a_h)} \cdot \\
&\quad \sum_{t=1}^{T} \underbrace{\left[\left(H-h+1-\sum_{h'=h}^{H} r_{h'}^{t,(h)}\right)\mathbf{1}\left\{(x_h^{t,(h)},a_h^{t,(h)}) = (x_h,a_h)\right\} - \mu_{1:h}^{\star,h}(x_h,a_h)L_h^t(x_h,a_h)\right]}_{:=\Delta_t^{x_h,a_h}},
\end{aligned} \tag{31}$$

where the last equality used the definition of the loss estimator $\widetilde{L}_h^t(x_h,a_h)$ in (12).

Observe that the random variables $\Delta_t^{x_h,a_h}$ satisfy the following:

- $\Delta_t^{x_h,a_h} \le H$ almost surely.
- $\mathbb{E}[\Delta_t^{(x_h,a_h)}|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information after iteration $t-1$. This follows as the episode was sampled using $\mu^{t,(h)} = \mu_{1:h}^{\star,h}\mu_{h+1:H}^t$, as well as the definition of $L_h^t(x_h,a_h)$ in (10).
- The conditional variance $\mathbb{E}[(\Delta_t^{(x_h,a_h)})^2|\mathcal{F}_{t-1}]$ can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left(\Delta_t^{(x_h,a_h)}\right)^2\middle|\mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\left(H-h+1-\sum_{h'=h}^{H} r_{h'}^{t,(h)}\right)^2 \mathbf{1}\left\{(x_h^{t,(h)},a_h^{t,(h)}) = (x_h,a_h)\right\}\middle|\mathcal{F}_{t-1}\right] \\
&\le H^2 \mathbb{P}^{\mu_{1:h}^{\star,h},\nu^t}\left((x_h^{t,(h)},a_h^{t,(h)}) = (x_h,a_h)\right) \\
&= H^2 \mu_{1:h}^{\star,h}(x_h,a_h) \cdot p_{1:h}^{\nu^t}(x_h).
\end{aligned}$$

Therefore, we can apply Freedman's inequality (Lemma B.1) and union bound to get that, for any fixed $\lambda \in (0, 1/H]$, with probability at least $1 - \delta/10$, the following holds simultaneously for all (h, x_h, a_h):

$$\sum_{t=1}^{T} \Delta_t^{(x_h,a_h)} \leq \lambda H^2 \mu_{1:h}^{\star,h}(x_h, a_h) \cdot \sum_{t=1}^{T} p_{1:h}^{\nu^t}(x_h) + \frac{\iota}{\lambda},$$

where $\iota := \log(10XA/\delta)$ is a log factor. Plugging this bound into (31) yields that, for all $h \in [H]$,

$$\begin{aligned}
&\mathrm{BIAS}_h^2 = \max_{\mu \in \Pi_{\max}} \sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{\mu_{1:h}(x_h, a_h)}{\mu_{1:h}^{\star,h}(x_h, a_h)} \cdot \sum_{t=1}^{T} \Delta_t^{x_h,a_h} \\
&\leq \max_{\mu \in \Pi_{\max}} \sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{\mu_{1:h}(x_h, a_h)}{\mu_{1:h}^{\star,h}(x_h, a_h)} \cdot \left[\lambda H^2 \mu_{1:h}^{\star,h}(x_h, a_h) \cdot \sum_{t=1}^{T} p_{1:h}^{\nu^t}(x_h) + \frac{\iota}{\lambda} \right] \\
&\leq \lambda H^2 \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h) \sum_{t=1}^{T} p_{1:h}^{\nu^t}(x_h) + \frac{\iota}{\lambda} \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{\mu_{1:h}(x_h, a_h)}{\mu_{1:h}^{\star,h}(x_h, a_h)} \\
&\overset{(i)}{=} \lambda H^2 T + \frac{\iota}{\lambda} \cdot X_h A.
\end{aligned}$$

Above, (i) used the fact that $\sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:h}(x_h, a_h) p_{1:h}^{\nu^t}(x_h) = 1$ for any $\mu \in \Pi_{\max}$ and any $t \in [T]$ (Lemma D.1(a)), as well as the balancing property of $\mu_{1:h}^{\star,h}$ (Lemma D.4). Combining the bounds for all $h \in [H]$, we get that with probability at least $1 - \delta/10$,

$$\sum_{h=1}^{H} \mathrm{BIAS}_h^2 \leq \lambda H^3 T + \frac{XA\iota}{\lambda}.$$

Choosing

$$\lambda = \min\left\{ \sqrt{\frac{XA\iota}{H^3 T}}, \frac{1}{H} \right\} \leq \frac{1}{H},$$

we obtain the bound

$$\sum_{h=1}^{H} \mathrm{BIAS}_h^2 \leq 2\sqrt{H^3 XAT\iota} + HXA\iota.$$

This is the desired result. □

G.5 Proof of Lemma G.4

Recall that for all (h, x_h), we have implemented Line 8 of Algorithm 2 as the Hedge algorithm (Algorithm 3) with learning rate $\eta \mu_{1:h}^{\star,h}(x_h, a)$ and loss vector $\left\{ \widetilde{L}_h^t(x_h, a) \right\}_{a \in \mathcal{A}}$ (cf. (11)). Therefore,

applying the standard regret bound for HEDGE (Lemma C.1), we get (below $a \in \mathcal{A}$ is arbitrary)

$$
\begin{aligned}
&\widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h) = \max_{\mu_h^\dagger(\cdot|x_h)} \sum_{t=1}^T \left\langle \mu_h^t(\cdot|x_h) - \mu_h^\dagger(\cdot|x_h), \widetilde{L}_h^t(x_h,\cdot) \right\rangle \\
&\le \frac{\log A}{\eta \mu_{1:h}^{\star,h}(x_h,a)} + \frac{\eta}{2} \cdot \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu_{1:h}^{\star,h}(x_h,a_h) \cdot \mu_h^t(a_h|x_h) \Big(\widetilde{L}_h^t(x_h,a_h) \Big)^2 \\
&\overset{(i)}{=} \frac{\log A}{\eta \mu_{1:h}^{\star,h}(x_h,a)} \\
&\qquad + \frac{\eta}{2} \cdot \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu_{1:h}^{\star,h}(x_h,a_h) \mu_h^t(a_h|x_h) \cdot \frac{\Big(H - h + 1 - \sum_{h'=h}^H r_{h'}^{t,(h)}\Big)^2 \mathbf{1}\left\{ (x_h^{t,(h)}, a_h^{t,(h)}) = (x_h,a_h) \right\}}{\Big(\mu_{1:h}^{\star,h}(x_h,a_h)\Big)^2} \\
&\le \frac{\log A}{\eta \mu_{1:h}^{\star,h}(x_h,a)} + \frac{\eta H^2}{2} \cdot \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu_h^t(a_h|x_h) \cdot \frac{\mathbf{1}\left\{ (x_h^{t,(h)}, a_h^{t,(h)}) = (x_h,a_h) \right\}}{\mu_{1:h}^{\star,h}(x_h,a_h)}.
\end{aligned} \tag{32}
$$

Above, (i) used the form of $\widetilde{L}_h^t$ in (12). Plugging this into the definition of REGRET_h, we have

$$
\begin{aligned}
&\mathrm{REGRET}_h = \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h) \\
&\le \underbrace{\max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \cdot \frac{\log A}{\eta \mu_{1:h}^{\star,h}(x_h,a)}}_{\mathrm{I}_h} \\
&\qquad + \underbrace{\max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \cdot \frac{\eta H^2}{2} \cdot \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu_h^t(a_h|x_h) \cdot \frac{\mathbf{1}\left\{ (x_h^{t,(h)}, a_h^{t,(h)}) = (x_h,a_h) \right\}}{\mu_{1:h}^{\star,h}(x_h,a_h)}}_{\mathrm{II}_h}.
\end{aligned} \tag{33}
$$

We first calculate term I_h. We have

$$
\begin{aligned}
\mathrm{I}_h &\overset{(i)}{=} \frac{\log A}{\eta} \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{1}{A_h} \cdot \frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})}{\mu_{1:h}^{\star,h}(x_h,a_h)} \\
&= \frac{\log A}{\eta} \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h,a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{(\mu_{1:(h-1)} \mu_h^{\mathrm{unif}})(x_h,a_h)}{\mu_{1:h}^{\star,h}(x_h,a_h)} \\
&\overset{(ii)}{=} \frac{\log A}{\eta} \cdot X_h A = \frac{X_h A \log A}{\eta},
\end{aligned}
$$

where (i) follows by splitting the sum over a_h and using the fact that $\mu_{1:h}^{\star,h}(x_h,a)$ does not depend on a; (ii) follows from the balancing property of $\mu_{1:h}^{\star,h}$ (Lemma D.4).

Next, we bound term II_h. We have

$$
\begin{aligned}
&\mathrm{II}_h = \frac{\eta H^2}{2} \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1},a_{h-1}) \cdot \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu_h^t(a_h|x_h) \cdot \frac{\mathbf{1}\left\{ (x_h^{t,(h)}, a_h^{t,(h)}) = (x_h,a_h) \right\}}{\mu_{1:h}^{\star,h}(x_h,a_h)} \\
&= \frac{\eta H^2}{2} \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1},a_{h-1})}{\mu_{1:h}^{\star,h}(x_h,a)} \cdot \sum_{t=1}^T \underbrace{\sum_{a_h \in \mathcal{A}} \mu_h^t(a_h|x_h) \cdot \mathbf{1}\left\{ (x_h^{t,(h)}, a_h^{t,(h)}) = (x_h,a_h) \right\}}_{:= \overline{\Delta}_t^{x_h}}.
\end{aligned} \tag{34}
$$

The last equality above used the fact that $\mu^{\star,h}_{1:h}(x_h, a_h)$ does not depend on a_h (cf. (5)).

Observe that the random variables $\overline{\Delta}^{x_h}_t$ satisfy the following:

- $\overline{\Delta}^{x_h}_t \in [0,1]$ almost surely;
- $\mathbb{E}[\overline{\Delta}^{x_h}_t | \mathcal{F}_{t-1}] = \sum_{a_h \in \mathcal{A}} \mu^{\star,h}_{1:h}(x_h, a_h) \cdot \mu^t_h(a_h|x_h) p^{\nu^t}_{1:h}(x_h)$, where $\mathcal{F}_{t-1}$ is the σ-algebra containing all information after iteration $t-1$;
- The conditional variance $\mathrm{Var}[\overline{\Delta}^{x_h}_t | \mathcal{F}_{t-1}]$ can be bounded as

$$\begin{aligned}
&\mathrm{Var}\left[\overline{\Delta}^{x_h}_t \middle| \mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\left(\overline{\Delta}^{x_h}_t\right)^2 \middle| \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{a_h \in \mathcal{A}} \left(\mu^t_h(a_h|x_h)\right)^2 \mathbf{1}\left\{(x^{t,(h)}_h, a^{t,(h)}_h) = (x_h, a_h)\right\} \middle| \mathcal{F}_{t-1}\right] \\
&= \sum_{a_h \in \mathcal{A}} \left(\mu^t_h(a_h|x_h)\right)^2 \cdot \mathbb{P}^{\mu^{\star,h}_{1:h} \times \nu^t}\left((x^{t,(h)}_h, a^{t,(h)}_h) = (x_h, a_h)\right) \\
&= \sum_{a_h \in \mathcal{A}} \mu^{\star,h}_{1:h}(x_h, a_h) \cdot \left(\mu^t_h(a_h|x_h)\right)^2 \cdot p^{\nu^t}_{1:h}(x_h).
\end{aligned}$$

Therefore, we can apply Freedman's inequality (Lemma B.1) and a union bound to obtain that, for any $\lambda \in (0,1]$, with probability at least $1 - \delta/10$, the following holds simultaneously for all (h, x_h):

$$\begin{aligned}
&\sum_{t=1}^T \overline{\Delta}^{x_h}_t - \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu^{\star,h}_{1:h}(x_h, a_h) \cdot \mu^t_h(a_h|x_h) p^{\nu^t}_{1:h}(x_h) \\
&\le \lambda \cdot \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu^{\star,h}_{1:h}(x_h, a_h) \cdot \left(\mu^t_h(a_h|x_h)\right)^2 \cdot p^{\nu^t}_{1:h}(x_h) + \frac{\iota}{\lambda},
\end{aligned}$$

where $\iota := \log(10X/\delta)$ is a log factor. Plugging this bound into (34) yields that, for all $h \in [H]$,

$$\begin{aligned}
\mathrm{II}_h &\le \frac{\eta H^2}{2} \cdot \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1}, a_{h-1})}{\mu^{\star,h}_{1:h}(x_h, a)} \cdot \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu^{\star,h}_{1:h}(x_h, a_h) \cdot \mu^t_h(a_h|x_h) p^{\nu^t}_{1:h}(x_h) \\
&\quad + \frac{\eta H^2}{2} \cdot \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1}, a_{h-1})}{\mu^{\star,h}_{1:h}(x_h, a)} \\
&\qquad\qquad \cdot \left[\lambda \sum_{t=1}^T \sum_{a_h \in \mathcal{A}} \mu^{\star,h}_{1:h}(x_h, a_h) \cdot \left(\mu^t_h(a_h|x_h)\right)^2 \cdot p^{\nu^t}_{1:h}(x_h) + \frac{\iota}{\lambda}\right] \\
&\overset{(i)}{\le} \frac{\eta H^2}{2} \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:(h-1)}(x_{h-1}, a_{h-1}) \cdot \sum_{t=1}^T \mu^t_h(a_h|x_h) p^{\nu^t}_{1:h}(x_h) \\
&\quad + \frac{\eta H^2}{2} \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:(h-1)}(x_{h-1}, a_{h-1}) \cdot \lambda \sum_{t=1}^T \left(\mu^t_h(a_h|x_h)\right)^2 \cdot p^{\nu^t}_{1:h}(x_h) \\
&\quad + \frac{\eta H^2}{2} \cdot \frac{\iota}{\lambda} \cdot \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \frac{\mu_{1:(h-1)}(x_{h-1}, a_{h-1})}{\mu^{\star,h}_{1:h}(x_h, a)} \\
&\overset{(ii)}{\le} \frac{\eta H^2}{2}(1+\lambda) \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \mu_{1:(h-1)}(x_{h-1}, a_{h-1}) \cdot \sum_{t=1}^T \mu^t_h(a_h|x_h) p^{\nu^t}_{1:h}(x_h) \\
&\quad + \frac{\eta H^2}{2} \cdot \frac{\iota}{\lambda} \cdot \max_{\mu \in \Pi_{\max}} \sum_{(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}} \frac{(\mu_{1:(h-1)} \mu^{\mathrm{unif}}_h)(x_h, a_h)}{\mu^{\star,h}_{1:h}(x_h, a_h)} \\
&\overset{(iii)}{=} \frac{\eta H^2}{2}(1+\lambda) T + \frac{\eta H^2}{2} \cdot \frac{\iota}{\lambda} \cdot X_h A.
\end{aligned}$$

Above, (i) used again the fact that $\mu^{\star,h}_{1:h}(x_h, a) = \mu^{\star,h}_{1:h}(x_h, a_h)$ for any $a, a_h \in \mathcal{A}$; (ii) used the fact that $\mu^t_h(a_h|x_h) \le 1$; (iii) used the fact that $\sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}}(\mu_{1:(h-1)}\mu^t_h)(x_h, a_h)\nu^t(x_h) = 1$ for any $\mu \in \Pi_{\max}$ and any $t \in [T]$ (Lemma D.1(a)), as well as the balancing property of $\mu^{\star,h}_{1:h}$ (Lemma D.4).

Combining the bounds for I_h and II_h, we obtain that

$$\begin{aligned}
&\sum_{h=1}^{H}\mathrm{REGRET}_h \le \sum_{h=1}^{H}(\mathrm{I}_h + \mathrm{II}_h)\\
&\le \sum_{h=1}^{H}\left[\frac{X_h A\log A}{\eta} + \frac{\eta H^2}{2}(1+\lambda)T + \frac{\eta H^2 X_h A\iota}{2\lambda}\right]\\
&\le \frac{XA\iota}{\eta} + \frac{\eta H^3}{2}T + \frac{\eta H^2}{2}\left[\lambda\cdot HT + \frac{XA\iota}{\lambda}\right],
\end{aligned}$$

where we have redefined the log factor $\iota := \log(10XA/\delta)$. Choosing $\lambda = 1$, the above can be upper bounded by

$$\frac{XA\iota}{\eta} + \eta H^3 T + \frac{\eta H^2 XA\iota}{2}.$$

Further choosing $\eta = \sqrt{XA\iota/(H^3T)}$, we obtain the bound

$$\sum_{h=1}^{H}\mathrm{REGRET}_h \le 2\sqrt{H^3XAT\iota} + \sqrt{HX^3A^3\iota^3/(4T)}.$$

This is the desired result. □

H BALANCED CFR WITH REGRET MATCHING

Many real-world applications of CFR-type algorithms use Regret Matching (Hart & Mas-Colell, 2000) instead of Hedge as the regret minimizer, due to its practical advantages such as learning-rate free and pruning effects (Tammelin, 2014; Burch et al., 2019). In this section, we show that Balanced CFR instantiated with Regret Matching enjoys $\widetilde{\mathcal{O}}(\sqrt{H^3XA^2T})$ "regret" and $\widetilde{\mathcal{O}}((H^4(XA^2 + YB^2)/\varepsilon^2)$ sample complexity for learning ε-NE (Theorem H.1 & Corollary H.2). The sample complexity is also sharp in X, Y, though is A (or B) times worse than the Hedge version, which is expected due to the difference between the regret minimizers.

We consider instantiating Line 8 of Algorithm 2 using the following Regret Matching algorithm:

$$\begin{aligned}
&\mu^{t+1}_h(a|x_h) = \frac{\left[R^t_{x_h}(a)\right]_+}{\sum_{a'\in\mathcal{A}}\left[R^t_{x_h}(a')\right]_+},\\
&\text{where } R^t_{x_h}(a) := \sum_{\tau=1}^{t}\left\langle \mu^\tau_h(\cdot|x_h), \widetilde{L}^\tau_h(x_h,\cdot)\right\rangle - \widetilde{L}^\tau_h(x_h, a) \quad \text{for all } a\in\mathcal{A}.
\end{aligned} \tag{35}$$

We now present the main theoretical guarantees for Balanced CFR with regret matching. The proof of Theorem H.1 can be found in Section H.1.

Theorem H.1 ("Regret" bound for Balanced CFR with Regret Matching)**.** *Suppose the max player plays Algorithm 2 where each R_{x_h} is instantiated as the Regret Matching algorithm (35). Then the policies μ^t achieve the following regret bound with probability at least $1-\delta$:*

$$\widetilde{\mathfrak{R}}^T := \max_{\mu^\dagger\in\Pi_{\max}}\sum_{t=1}^{T}\left\langle \mu^t - \mu^\dagger, \ell^t\right\rangle \le \mathcal{O}(\sqrt{H^3XA^2T\iota}),$$

where $\iota = \log(10XA/\delta)$ is a log factor. Further, each round plays H episodes against ν^t (so that the total number of episodes played is HT).

We then have the following corollary directly by the regret-to-Nash conversion (Proposition 1).

Corollary H.2 (Learning Nash using Balanced CFR with Regret Matching)**.** *Letting both players play Algorithm 2 in a self-play fashion against each other for T rounds, where each R_{x_h} is instantiated as the Regret Matching algorithm (35). Then, for any $\varepsilon > 0$, the average policy $(\overline{\mu}, \overline{\nu}) = (\frac{1}{T}\sum_{t=1}^{T}\mu^t, \frac{1}{T}\sum_{t=1}^{T}\nu^t)$ achieves $\mathrm{NEGap}(\overline{\mu}, \overline{\nu}) \le \varepsilon$ with probability at least $1-\delta$, as long as*

$$T \ge \mathcal{O}(H^3(XA^2 + YB^2)\iota/\varepsilon^2),$$

where $\iota := \log(10(XA+YB)/\delta)$ is a log factor. The total amount of episodes played is at most

$$2H \cdot T = \mathcal{O}(H^4(XA^2 + YB^2)\iota/\varepsilon^2).$$

H.1 Proof of Theorem H.1

The proof is similar as Theorem 6, except for plugging in the regret bound for Regret Matching instead of Hedge.

First, by Lemma G.1, we have $\widetilde{\mathfrak{R}}^T \le \sum_{h=1}^{H}\mathfrak{R}_h^T$, where for any $h \in [H]$ we have

$$\begin{aligned}
\mathfrak{R}_h^T \le& \underbrace{\max_{\mu\in\Pi_{\max}} \sum_{x_h\in\mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1}, a_{h-1})\widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h)}_{:=\mathrm{REGRET}_h} \\
&+ \underbrace{\max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:(h-1)}(x_{h-1}, a_{h-1}) \sum_{t=1}^{T} \mu_h^t(a_h|x_h)\Big[L_h^t(x_h,a_h) - \widetilde{L}_h^t(x_h,a_h)\Big]}_{:=\mathrm{BIAS}_h^1} \\
&+ \underbrace{\max_{\mu\in\Pi_{\max}} \sum_{(x_h,a_h)\in\mathcal{X}_h\times\mathcal{A}} \mu_{1:h}(x_h, a_h) \sum_{t=1}^{T} \Big[\widetilde{L}_h^t(x_h,a_h) - L_h^t(x_h,a_h)\Big]}_{:=\mathrm{BIAS}_h^2} \\
=& \mathrm{REGRET}_h + \mathrm{BIAS}_h^1 + \mathrm{BIAS}_h^2,
\end{aligned} \tag{36}$$

where the definition of $\widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h)$, $L_h^t(x_h,a_h)$ are at the beginning of Section G.1 and the definition of $\widetilde{L}_h^t(x_h,a_h)$ are given by Algorithm 2.

To upper bound BIAS_h^1 and BIAS_h^2, we use the same strategy as the proof of Lemma G.2 and G.3 (whose proofs are independent of the regret minimizer), so that we have the same bound as in Lemma G.2 and G.3: with probability at least $1-\delta/5$, we have

$$\sum_{h=1}^{H}\mathrm{BIAS}_h^1 \le 2\sqrt{H^3XAT\iota} + HX\iota, \quad \sum_{h=1}^{H}\mathrm{BIAS}_h^2 \le 2\sqrt{H^3XAT\iota} + HXA\iota, \tag{37}$$

where $\iota = \log(10XA/\delta)$.

To upper bound REGRET_h, we use the same strategy as the proof of Lemma G.4 as in Section G.5. First, applying the regret bound for Regret Matching (Lemma C.2 & Remark C.3), we get (below $a \in \mathcal{A}$ is arbitrary, and $\eta > 0$ is also arbitrary)

$$\begin{aligned}
\widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h) =& \max_{\mu_h^\dagger(\cdot|x_h)} \sum_{t=1}^{T} \Big\langle \mu_h^t(\cdot|x_h) - \mu_h^\dagger(\cdot|x_h), \widetilde{L}_h^t(x_h,\cdot)\Big\rangle \\
\le& \frac{1}{\eta\mu_{1:h}^{\star,h}(x_h,a)} + \frac{\eta}{2}\cdot\sum_{t=1}^{T}\sum_{a_h\in\mathcal{A}} \mu_{1:h}^{\star,h}(x_h,a_h)\cdot A\bar{\mu}_h^t(a_h|x_h)\Big(\widetilde{L}_h^t(x_h,a_h)\Big)^2 \\
\le& \frac{1}{\eta\mu_{1:h}^{\star,h}(x_h,a)} + \frac{\eta H^2}{2}\cdot\sum_{t=1}^{T}\sum_{a_h\in\mathcal{A}} A\cdot\bar{\mu}_h^t(a_h|x_h)\cdot\frac{\mathbf{1}\Big\{(x_h^{t,(h)}, a_h^{t,(h)}) = (x_h,a_h)\Big\}}{\mu_{1:h}^{\star,h}(x_h,a_h)},
\end{aligned} \tag{38}$$

where $\bar{\mu}_h^t(a_h|x_h) = (\mu_h^t(a_h|x_h) + (1/A))/2$ is a probability distribution over $[A]$. Comparing the right hand side of Eq. (38) with the right hand side of Eq. (32), we can see that there is only one difference which is $A \cdot \bar{\mu}_h^t$ versus μ_h^t. Plugging this into the definition of REGRET_h, we have

$$\begin{aligned}
&\mathrm{REGRET}_h = \max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1}, a_{h-1}) \widetilde{\mathfrak{R}}_h^{\mathrm{imm},T}(x_h) \\
&\leq \underbrace{\max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1}, a_{h-1}) \cdot \frac{1}{\eta \mu_{1:h}^{\star,h}(x_h, a)}}_{\mathrm{I}_h} \\
&\quad + \underbrace{\max_{\mu \in \Pi_{\max}} \sum_{x_h \in \mathcal{X}_h} \mu_{1:(h-1)}(x_{h-1}, a_{h-1}) \cdot \frac{\eta H^2}{2} \cdot \sum_{t=1}^{T} \sum_{a_h \in \mathcal{A}} A \cdot \bar{\mu}_h^t(a_h|x_h) \cdot \frac{\mathbf{1}\left\{(x_h^{t,(h)}, a_h^{t,(h)}) = (x_h, a_h)\right\}}{\mu_{1:h}^{\star,h}(x_h, a_h)}}_{\mathrm{II}_h}.
\end{aligned} \tag{39}$$

Comparing Eq. (39) with Eq. (33), we can see that I_h in Eq. (39) is the same as I_h in Eq. (33), and II_h in Eq. (39) and (33) only have one difference which is also $A \cdot \bar{\mu}_h^t$ versus μ_h^t. Using the same argument as in the former proof, we have

$$\mathrm{I}_h = \frac{X_h A}{\eta}.$$

Furthermore, using the same argument as in the former proof, we can show that the upper bound of II_h in Eq. (39) is at most A times the upper bound of II_h in Eq. (33). This gives for any $\lambda \in (0, 1)$, with probability at least $1 - \delta/10$, we have

$$\mathrm{II}_h \leq \frac{\eta H^2 A}{2}(1 + \lambda)T + \frac{\eta H^2}{2} \cdot \frac{\iota}{\lambda} \cdot X_h A^2.$$

Combining the bounds for I_h and II_h, we obtain that

$$\sum_{h=1}^{H} \mathrm{REGRET}_h \leq \sum_{h=1}^{H} (\mathrm{I}_h + \mathrm{II}_h) \leq \frac{XA}{\eta} + \frac{\eta H^3 A}{2} T + \frac{\eta H^2 A}{2}\left[\lambda \cdot HT + \frac{XA\iota}{\lambda}\right],$$

Choosing $\lambda = 1$ and choosing $\eta = \sqrt{X\iota/(H^3 T)}$, with probability at least $1 - \delta/10$, we obtain the bound

$$\sum_{h=1}^{H} \mathrm{REGRET}_h \leq 2\sqrt{H^3 X A^2 T \iota} + \sqrt{H X^3 A^4 \iota^3/(4T)}. \tag{40}$$

This bound is $\sqrt{A}$ times larger than the bound of $\sum_{h=1}^{H} \mathrm{REGRET}_h$ as in Lemma G.4.

Combining Eq. (36), (37) and (40), we obtain the following: with probability at least $1 - 3\delta/10 \geq 1 - \delta$, we have

$$\begin{aligned}
\widetilde{\mathfrak{R}}^T &\leq \sum_{h=1}^{H} \mathfrak{R}_h^T \leq \sum_{h=1}^{H} \mathrm{REGRET}_h + \sum_{h=1}^{H} \mathrm{BIAS}_h^1 + \sum_{h=1}^{H} \mathrm{BIAS}_h^2 \\
&\leq 6\sqrt{H^3 X A^2 T \iota} + 2HXA\iota + \sqrt{H X^3 A^4 \iota^3/(4T)}.
\end{aligned}$$

Additionally, recall the naive bound $\widetilde{\mathfrak{R}}^T \leq HT$ on the regret (which follows as $\langle \mu^t, \ell^t \rangle \in [0, H]$ for any $\mu \in \Pi_{\max}$, $t \in [T]$), we get

$$\begin{aligned}
\widetilde{\mathfrak{R}}^T &\leq \min\left\{6\sqrt{H^3 X A^2 T \iota} + 2HXA\iota + \sqrt{H X^3 A^4 \iota^3/4T}, HT\right\} \\
&\leq HT \cdot \min\left\{6\sqrt{HXA^2\iota/T} + 2XA\iota/T + \sqrt{X^3 A^4 \iota^3/(4HT^3)}, 1\right\}.
\end{aligned}$$

For $T > HXA^2\iota$, the min above is upper bounded by $9\sqrt{HXA^2\iota/T}$. For $T \leq HXA^2\iota$, the min above is upper bounded by $1 \leq 9\sqrt{HXA^2\iota/T}$. Therefore, we always have

$$\widetilde{\mathfrak{R}}^T \leq HT \cdot 9\sqrt{HXA^2\iota/T} = 9\sqrt{H^3 X A^2 T \iota}.$$

This is the desired result. □

I EXTENSION TO MULTI-PLAYER GAMES

In this section, we show that our Balanced OMD and Balanced CFR generalize directly to learning Coarse Correlated Equilibria in multi-player general-sum games.

I.1 DEFINITION OF MULTI-PLAYER GENERAL-SUM IIEFGS

We consider an m-player general-sum IIEFG with X_i infosets and A_i actions for the i-th player. Our definition is parallel to the POMG formulation for two-player zero-sum IIEFGs in Section 2.

Partially observable Markov games We consider finite-horizon, tabular, m-player general-sum Markov Games with partial observability. Formally, it can be described as a $\mathrm{POMG}(H, \mathcal{S}, \{\mathcal{X}_i\}_{i=1}^m, \{\mathcal{A}_i\}_{i=1}^m, \mathbb{P}, \{r_i\}_{i=1}^m)$, where

- H is the horizon length;
- $\mathcal{S} = \bigcup_{h\in[H]} \mathcal{S}_h$ is the (underlying) state space;
- $\mathcal{X}_i = \bigcup_{h\in[H]} \mathcal{X}_{i,h}$ is the space of infosets for the *i-th player* with $|\mathcal{X}_{i,h}| = X_{i,h}$ and $X_i := \sum_{h=1}^H X_{i,h}$. At any state $s_h \in \mathcal{S}_h$, the i-th player only observes the infoset $x_{i,h} = x_i(s_h) \in \mathcal{X}_{i,h}$, where $x_i : \mathcal{S} \to \mathcal{X}_i$ is the emission function for the i-th player;
- $\mathcal{A}_i$ is the action spaces for the i-th player with $|\mathcal{A}_i| = A_i$. For any h, we define the joint action of m players by $\mathbf{a}_h := (a_{1,h}, \cdots, a_{m,h})$ and the set of joint actions by $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$.
- $\mathbb{P} = \{p_1(\cdot) \in \Delta(\mathcal{S}_1)\} \cup \{p_h(\cdot|s_h, \mathbf{a}_h) \in \Delta(\mathcal{S}_{h+1})\}_{(s_h,\mathbf{a}_h)\in\mathcal{S}_h\times\mathcal{A},\ h\in[H-1]}$ are the transition probabilities, where $p_1(s_1)$ is the probability of the initial state being s_1, and $p_h(s_{h+1}|s_h, \mathbf{a}_h)$ is the probability of transitting to s_{h+1} given state-action (s_h, a_h, b_h) at step h;
- $r_i = \{r_{i,h}(s_h, \mathbf{a}_h) \in [0, 1]\}_{(s_h,\mathbf{a}_h)\in\mathcal{S}_h\times\mathcal{A}}$ are the (random) reward functions with mean $\overline{r}_{i,h}(s_h, \mathbf{a}_h)$.

Policies, value functions A policy for the i-th player is denoted by $\pi_i = \{\pi_{i,h}(\cdot|x_{i,h}) \in \Delta(\mathcal{A}_i)\}_{h\in[H],x_{i,h}\in\mathcal{X}_{i,h}}$, where $\pi_{i,h}(a_{i,h}|x_{i,h})$ is the probability of taking action $a_{i,h} \in \mathcal{A}_i$ at infoset $x_{i,h} \in \mathcal{X}_{i,h}$. A trajectory for the i-th player takes the form $(x_{i,1}, a_{i,1}, r_{i,1}, x_{i,2}, \ldots, x_{i,H}, a_{i,H}, r_{i,H})$, where $a_{i,h} \sim \pi_{i,h}(\cdot|x_{i,h})$, which depends on both the other (unseen) players' policy and underlying state transition.

We use π to denote the joint policy. Notice although the marginals are π_i, π is not necessarily a product policy. When π is indeed a product policy, we have $\pi = \pi_1 \times \cdots \times \pi_m$. We also use π_{-i} to denote the joint product policy excluding the i-th player. The overall game value of the i-th player for any joint policy π is denoted by $V_i^\pi := \mathbb{E}_\pi\left[\sum_{h=1}^H r_{i,h}(s_h, \mathbf{a}_h)\right]$.

Tree structure and perfect recall As before, we assume

- *Tree structure*: for any h and $s_h \in \mathcal{S}_h$, there exists a unique history $(s_1, \mathbf{a}_1, \ldots, s_{h-1}, \mathbf{a}_{h-1})$ of past states and (joint) actions that leads to h.
- *Perfect recall*: For any h and any infoset $x_{i,h} \in \mathcal{X}_{i,h}$ for the i-th player, there exists a unique history $(x_{i,1}, a_{i,1}, \ldots, x_{i,h-1}, a_{i,h-1})$ of past infosets and i-th player's actions that leads to $x_{i,h}$.

Given above conditions, under any product policy π, the probability of reaching state-action $(s_h, \mathbf{a}_h)$ at step h takes the form

$$\mathbb{P}^\pi(s_h, \mathbf{a}_h) = p_{1:h}(s_h) \prod_{i=1}^m \pi_{i,1:h}\left(x_{i,h}, a_{i,h}\right), \tag{41}$$

where $\{s_{h'}, \mathbf{a}_{h'}\}_{h'\le h-1}$ are the histories uniquely determined from s_h and $x_{i,h'} = x_i(s_{h'})$. We have also defined the sequence-form transition probability as

$$p_{1:h}(s_h) := p_1(s_1) \prod_{h\prime\le h-1} p_{h\prime}(s_{h\prime+1}|s_{h\prime}, \mathbf{a}_{h\prime}),$$

and the *sequence-form policies* as

$$\pi_{i,1:h}\left(x_{i,h}, a_{i,h}\right) := \prod_{h\prime=1}^{h} \pi_{i,h'}\left(a_{i,h'}|x_{i,h'}\right).$$

Regret and CCE Similar as how regret minimization in two-player zero-sum games leads to an approximate Nash equilibrium (Proposition 1), in multi-player general-sum games, regret minimization is known to lead to an approximate NFCCE. Let $\{\pi^t\}_{t=1}^T$ denote a sequence of joint policies (for all players) over T rounds. The regret of the i-th player is defined by

$$\mathfrak{R}_i^T := \max_{\pi_i^\dagger \in \Pi_i} \sum_{t=1}^{T} \left(V_i^{\pi_i^\dagger, \pi_{-i}^t} - V_i^{\pi^t} \right).$$

where Π_i denotes the set of all possible policies for the i-th player.

Definition I.1 (NFCCE)**.** *A joint policy π (for all players) is an ε-approximate Normal-Form Coarse Correlated Equilibrium (NFCCE) if*

$$\text{CCEGap}(\pi) := \max_{i\in[m]} \Big(\max_{\pi_i^\dagger \in \Pi_i} V_i^{\pi_i^\dagger, \pi_{-i}} - V_i^{\pi} \Big) \leq \varepsilon,$$

i.e., no player can gain more than ε in her own reward by deviating from π and playing some other policy on her own.

Using online-to-batch conversion, it is a standard result that sub-linear regret for all the players ensures that the average policy $\overline{\pi}$ is an approximate NFCCE (Celli et al., 2019).

Proposition I.2 (Regret-to-CCE conversion for multi-player general-sum games)**.** *Let the average policy $\overline{\pi}$ be defined as playing a policy within $\{\pi^t\}_{t=1}^T$ uniformly at random, then we have*

$$\text{CCEGap}(\overline{\pi}) = \frac{\max_{i\in[m]} \mathfrak{R}_i^T}{T}.$$

We include a short justification for this standard result here for completeness.

Proof. By definition of $\overline{\pi}$, we have for any $i \in [m]$ and $\pi_i^\dagger \in \Pi_i$ that

$$V_i^{\pi_i^\dagger, \overline{\pi}_{-i}} - V_i^{\overline{\pi}} = \frac{1}{T} \sum_{t=1}^{T} \left(V_i^{\pi_i^\dagger, \pi_{-i}^t} - V_i^{\pi^t} \right).$$

Taking the max over $\pi_i^\dagger \in \Pi_i$ and $i \in [m]$ on both sides yields the desired result. □

We remark that the NFCCE differs from other types of Coarse Correlated Equilibria in the literature such as the EFCCE[5] (Farina et al., 2020a).

I.2 Learning NFCCE in multi-player IIEFGs

Using the known connection between no-regret and NFCCE (Celli et al., 2019), we can learn an ε-NFCCE in an multi-player IIEFG sample-efficiently by letting all players run either Balanced CFR or Balanced OMD in a self-play fashion. In the following, we let $\{\pi_i^t\}_{t=1}^T$ denote the policies maintained by player i, and $\pi^t := \prod_{i=1}^m \pi_i^t$ denote their joint policy in the t-th round.

Theorem I.3 (Learning NFCCE sample-efficiently using Balanced OMD / Balanced CFR)**.** *We have*

(a) (Balanced OMD) Let all players play Algorithm 1 for T rounds with learning rate $\eta = \sqrt{X_i A_i \log A_i/(H^3 T)}$ and IX parameter $\gamma = \sqrt{X_i A_i \iota/(HT)}$ for the i-th player. Then for any $\varepsilon > 0$, the average policy $\overline{\pi}$ uniformly sampled from $\{\pi^t\}_{t=1}^T$ satisfies $\text{CCEGap}(\overline{\pi}) \leq \varepsilon$ with probability at least $1-\delta$, as long as the number of episodes

$$T \geq \mathcal{O}\Big(H^3 \iota \Big(\max_{i\in[m]} X_i A_i \Big) / \varepsilon^2 \Big),$$

where $\iota := \log(3H \sum_{i=1}^m X_i A_i / \delta)$ is a log factor.

[5] Such distinctions only exist for (Coarse) Correlated Equilibria and not for the NE studied in the previous sections.

(b) (Balanced CFR) Let all players play Algorithm 2 in the same self-play fashion as Corollary 7 for T rounds, with R_{x_h} instantiated as Hedge (11) with learning rate $\eta = \sqrt{X_i A_i \iota/(H^3 T)}$ for the i-th player. Then for any $\varepsilon > 0$, the average policy $\overline{\pi}$ uniformly sampled from $\{\pi^t\}_{t=1}^T$ satisfies $\mathrm{CCEGap}(\overline{\pi}) \le \varepsilon$ with probability at least $1-\delta$, as long as $T \ge \mathcal{O}\big(H^3 \iota (\max_{i\in[m]} X_i A_i)/\varepsilon^2\big)$. The total number of episodes played is at most

$$mH \cdot T = \mathcal{O}\Big(H^4 m\iota \cdot \Big(\max_{i\in[m]} X_i A_i\Big)/\varepsilon^2\Big).$$

where $\iota := \log(10 \sum_{i=1}^m X_i A_i/\delta)$ is a log factor.

For both algorithms, the number of episodes for learning an ε-NFCCE scales linearly with $\max_{i\in[m]} X_i A_i$ (with Balanced CFR having an additional Hm factor than Balanced OMD), compared to the best existing $\max_{i\in[m]} X_i^2 A_i$ dependence (e.g. by self-playing IXOMD (Kozuno et al., 2021)).

Proof. It is straightforward to see that the regret guarantees for Balanced OMD (Theorem 4) and Balanced CFR (Theorem 6) also hold in multi-player general-sum games (e.g. by modeling all other players as a single opponent). Therefore, the regret-to-CCE conversion in Proposition I.2 directly implies that, letting $\overline{\pi}$ denote the joint policy of playing a uniformly sampled policy within $\{\pi^t\}_{t=1}^T$, we have for Balanced OMD that

$$\mathrm{CCEGap}(\overline{\pi}) \le \mathcal{O}\left(\frac{\max_{i\in[m]} \sqrt{H^3 X_i A_i \iota T}}{T}\right) = \mathcal{O}\left(\sqrt{H^3 \Big(\max_{i\in[m]} X_i A_i\Big) \iota/T}\right),$$

with probability at least $1-\delta$, where $\iota := \log(3H \sum_{i=1}^m X_i A_i/\delta)$ is a log factor. Choosing $T \ge \widetilde{\mathcal{O}}\big(H^3 \big(\max_{i\in[m]} X_i A_i\big) \iota/\varepsilon^2\big)$ ensures that the right-hand side is at most ε. This shows part (a). A similar argument can be done for the Balanced CFR algorithm to show part (b). □