Adapting to game trees in zero-sum imperfect information games

Côme Fiegel [1]   Pierre Ménard [2]   Tadashi Kozuno [3]   Rémi Munos [4]   Vianney Perchet [1 5]   Michal Valko [4]

Abstract

Imperfect information games (IIG) are games in which each player only partially observes the current game state. We study how to learn ε-optimal strategies in a zero-sum IIG through self-play with *trajectory feedback*. We give a problem-independent lower bound $\widetilde{\mathcal{O}}(H(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$ on the required number of realizations to learn these strategies with high probability, where H is the length of the game, $A_{\mathcal{X}}$ and $B_{\mathcal{Y}}$ are the total number of actions for the two players. We also propose two Follow the Regularized leader (FTRL) algorithms for this setting: `Balanced FTRL` which matches this lower bound, but requires the knowledge of the information set structure beforehand to define the regularization; and `Adaptive FTRL` which needs $\widetilde{\mathcal{O}}(H^2(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$ realizations without this requirement by progressively adapting the regularization to the observations.

1. Introduction

In imperfect information games (IIG), players, upon taking an action, may only have access to *partial information about the true current game state*. This type of games allows for the modelling of complex strategic behavior such as bluffing (Koller & Pfeffer, 1995). In this work, in particular, we study extensive-form two-players zero-sum IIGs.

The *extensive-form* of a game describes, which player is playing and which actions are available, sequentially and depending on the previous moves. It is typically represented by a tree of depth H, where nodes are *states* that are controlled by one of the players that decides, based on each player action, the next game states among its children.

Players can be uncertain about the true game state upon playing. We therefore assume that the set of states controlled by the player is partitioned into *information sets*, where an information set is a set of states that are indistinguishable to this player. As commonly assumed in game theory, we suppose *perfect recall* (Kuhn, 1950): players remember their previous moves, which implies that the space of information sets has a *tree structure*.

We focus more specifically on *zero-sum games* (what the max-player gains is the opposite of the other, min-player) and aim at devising algorithms that learn ε-optimal strategies (von Neumann, 1928). For this purpose, we actually develop a unilateral algorithm that dictates to a single player how to play over multiple realizations of the game called *episodes*, each of the same length H, in order to minimize the *regret*—the difference between the cumulative gain and the gain of the best fixed strategy—against any sequence of moves of the adversary. If both players follows such an algorithm, we can then prove that their average strategy becomes more precise over time. We especially want to construct algorithms with a regret that scales as slowly as possible with respect to the different game parameters.

For any information set $x \in \mathcal{X}$, we define by A_x the number of available actions at x, as two indistinguishable states must have the same number of actions. The total number of actions of the max-player is then defined by $A_{\mathcal{X}} = \sum_{x \in \mathcal{X}} A_x$. Similarly, we let $\mathcal{Y}$ be the collection of information sets of the min-player, B_y the number of actions available at $y \in \mathcal{Y}$ and $B_{\mathcal{Y}} = \sum_{y \in \mathcal{Y}} B_y$ the total number of actions. Meanwhile X and Y denote the number of information sets for the max and min-players.

Related work   When the structure of the game, transition probabilities and reward function are known beforehand, several methods exist to approximate the optimal strategy. A common approach is to reformulate the problem as a linear program that can be solved efficiently (Romanovsky, 1962; von Stengel, 1996; Koller et al., 1996). Some methods see it as a saddle-point problem and rely on first-order optimization methods (Hoda et al., 2010; Kroer et al., 2015; 2018; 2020; Munos et al., 2020; Lee et al., 2021). Other common approaches try to locally minimize the regret at each information set, e.g., *counterfactual regret minimization* (Zinkevich et al., 2007; Tammelin, 2014; Burch et al., 2019).

[1]CREST, ENSAE, IP Paris, Palaiseau, France [2]ENS Lyon, Lyon, France [3]Omron Sinic X, Tokyo, Japan [4]Deepmind, Paris, France [5]CRITEO AI Lab, Paris, France. Correspondence to: Côme Fiegel <come.fiegel@normalesup.org>.

The main practical weakness of the above approaches is their prohibitive cost as the size of the game grows. Indeed, the computations are mostly done on the whole information set trees, which implies a time complexity of at least $\mathcal{O}(X+Y)$ at each step at best (Lanctot et al., 2009). All the recent successes in empirically solving large IIGs actually avoid the direct use of these full information algorithms (Moravčík et al., 2017; Brown & Sandholm, 2018; Schmid et al., 2021; Bakhtin et al., 2022; Pérolat et al., 2022). In addition, depending on the practical case, it may not be possible to know the state transitions, or even to start the game at an arbitrary state.

We therefore treat the settings when the algorithms can only access a *trajectory feedback* of the game: they only have access to the observations of the player. For this case, the outcome sampling MCCFR (Lanctot et al., 2009; Schmid et al., 2018; Farina et al., 2020) feeds estimates of the counterfactual regret to the CFR algorithm, which is by design a full-information feedback algorithm. Precisely, outcome-sampling MCCFR uses a uniform sampling of the actions in each information set, which is actually sub-optimal when the information set trees of the players are unbalanced, as information sets in smaller sub-trees are then observed way more times than information sets of bigger sub-trees. For this reason, Farina et al. (2020) instead propose to build an estimate of the counterfactual regret according to a *balanced policy* that, roughly speaking, samples actions proportionally to the size of the associated sub-trees.[1] While this balanced policy allows for a better sample complexity, it also requires to know the game structure beforehand.

Closer to the adversarial bandits, Farina et al. (2021b) propose to use the *online mirror descent* (OMD) algorithm along with a dilatation of the Shannon entropy on all information sets acting as a regularizer (Hoda et al., 2010). However, their approach uses importance sampling (Auer et al., 2003) to estimate the losses, whose variance is too large to allow the computation of ε-optimal strategies with high probability. Kozuno et al. (2021) solved this issue by using a biased estimate called *Implicit exploration* (IX, Kocák et al., 2014; Neu, 2015; Lattimore & Szepesvári, 2020), and proved that their IXOMD algorithm achieves a sample complexity of order $\widetilde{\mathcal{O}}(H^2(XA_{\mathcal{X}}+YB_{\mathcal{Y}})/\varepsilon^2)$ with high probability.[2] Unfortunately, it does not match the $\widetilde{\mathcal{O}}(H(A_{\mathcal{X}}+B_{\mathcal{Y}})/\varepsilon^2)$ lower bound, see Table 1.

More recently, Bai et al. (2022) propose to weight differently each of the information set in the regularizer using the aforementioned balanced policy. They obtain a sample com-

plexity of $\widetilde{\mathcal{O}}(H^3(A_{\mathcal{X}}+B_{\mathcal{Y}})/\varepsilon^2)$ that scales linearly with the total number of actions, but again, at the extra price of requiring the knowledge of the game structure beforehand.

Hence, we raise the following questions: *What is the optimal rate for learning ε-optimal strategies through self-play? Is it possible to learn these ε-optimal strategies with a complexity scaling linearly with the total number of actions without the prior knowledge of the information set tree structure?* In this work we answer *both questions*.

Contributions   We make the following contributions:

- We prove a lower bound of $\mathcal{O}(H(A_{\mathcal{X}}+B_{\mathcal{Y}})/\varepsilon^2)$ on the number of necessary plays required to compute an ε-optimal profile with only a trajectory feedback. Compared to the one stated by Bai et al. (2022), it applies to algorithms that are only correct with a high probability, and enjoys an extra H factor if the size of the action set is allowed to vary with the information set.

- We propose the `Balanced FTRL` algorithm that matches this lower bound, up to logarithm factors, using the knowledge of the structure of the information sets. It uses the Follow The Regularized Leader (FTRL) algorithm instead of OMD and follows the idea of balancing, but using instead a concept of *balanced transition*. In addition to the use of Shannon entropy, it also allows the use of the Tsallis entropy, in order to get a better sample complexity at the price of increased computation time.

- We then propose the `Adaptive FTRL` algorithm which matches this lower bound up to an extra H and logarithmic factors, *without* using the knowledge of the structure. Instead of using the balanced transition, it estimates the *actual* transitions of each episode.

- We provide a practical implementation of `Balanced FTRL` and `Adaptive FTRL` through Algorithm 3 of Appendix F. This implementation is computationally efficient, as the policy is only updated on the information sets visited in the previous episode.

- Finally, we provide experiments that compare the performances of these two algorithms with IXOMD and `Balanced OMD`. We observe that, despite having different theoretical guarantees, the algorithms all seem to have comparable performances in practice.

All rates are summarized in Table 1.

2. Setting

Consider an episodic, finite, two-player, zero-sum IIG $(\mathcal{S}, \mathcal{X}, \mathcal{Y}, \{\mathcal{A}(x)\}_{x\in\mathcal{X}}, \{\mathcal{B}(y)\}_{y\in\mathcal{Y}}, H, p, r)$, which consists of the following components (Kuhn, 1950; Littman, 1994):

- A finite state space $\mathcal{S}$ and two information set spaces (partitions of $\mathcal{S}$) $\mathcal{X}$ of size X and $\mathcal{Y}$ of size Y for the max- and min-player respectively.

---

[1]The exact definition of the balanced policy varies among different works, but the main idea remains the same.

[2]For algorithms with a probability at least $1-\delta$ of a correct output, the symbol $\widetilde{\mathcal{O}}$ hides dependencies logarithmic in $H, A_{\mathcal{X}}, B_{\mathcal{Y}}, \delta$ and ε.

Algorithm	Sample complexity	Structure-free
MCCFR (Farina et al., 2020; Bai et al., 2022)	$\widetilde{\mathcal{O}}(H^4(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$	✗
IXOMD (Kozuno et al., 2021)	$\widetilde{\mathcal{O}}(H^2(X A_{\mathcal{X}} + Y B_{\mathcal{Y}})/\varepsilon^2)$	✓
Balanced OMD (Bai et al., 2022)	$\widetilde{\mathcal{O}}(H^3(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$	✗
Balanced FTRL (this paper)	$\widetilde{\mathcal{O}}(H(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$	✗
Adaptive FTRL (this paper)	$\widetilde{\mathcal{O}}(H^2(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$	✓
Lower bound (this paper)	$\widetilde{\mathcal{O}}(H(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$	

Table 1. Sample complexity for episodic, finite, two-player, zero-sum IIGs. *Structure-free* algorithm designates an algorithm that does not need to know the structure of the information set spaces in advance. The symbol $\widetilde{\mathcal{O}}$ hides dependencies logarithmic in $A_{\mathcal{X}}, B_{\mathcal{Y}}, \varepsilon$ and δ. Note that for algorithms that only work with fixed action sets of size A and B we have $A_{\mathcal{X}} = AX$ and $B_{\mathcal{Y}} = BY$.

- For each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, finite action spaces $\mathcal{A}(x)$ of size A_x and $\mathcal{B}(y)$ of size B_y.
- The length $H \in \mathbb{N}$ of the game.
- Initial state distribution $p_0 \in \Delta(\mathcal{S})$ and a state-transition probability kernel $(p_h)_{h \in [H-1]}$ with[3] $p_h : \mathcal{S} \times \mathcal{A}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y}) \to \Delta(\mathcal{S})$ for each $h \in [H-1]$.
- A reward function $(r_h)_{h \in [H]}$ with $r_h : \mathcal{S} \times \mathcal{A}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y}) \to [0, 1]$.

Perfect-recall As explained in the introduction, we assume perfect-recall. Formally, this means that for the max-player and for each information set $x \in \mathcal{X}$ there exists a unique $h \in [H]$ and history $(x_1, a_1, ..., x_h)$ such that $x_h = x$. Specifically, we write $x \geq x'$ if x' is part of the history that leads to x. With this assumption, both $\mathcal{X}$ and $\mathcal{Y}$ can be partitioned into H different subsets $(\mathcal{X}_h)_{h \in [H]}$ and $(\mathcal{Y}_h)_{h \in [H]}$, $\mathcal{X}_h$ and $\mathcal{Y}_h$ being the sets of possible information sets at time step h for respectively the max and min-player.

For convenience we let $\mathcal{A}(\mathcal{X}_h) := \{(x_h, a_h) : x_h \in \mathcal{X}_h \text{ and } a_h \in \mathcal{A}(x_h)\}$ be the total action set for the max-player at depth h, and use an analogous notation $\mathcal{B}(\mathcal{Y}_h)$ for the min-player. The unions of these sets are denoted by $\mathcal{A}(\mathcal{X}) := \cup_{h \in [H]} \mathcal{A}(\mathcal{X}_h)$, $\mathcal{B}(\mathcal{Y}) := \cup_{h \in [H]} \mathcal{B}(\mathcal{Y}_h)$ with their respective sizes $A_{\mathcal{X}}$ and $B_{\mathcal{Y}}$, as defined in the introduction.

Policies The perfect-recall assumption allows us to represent a policy of the max-player as a sequence $\mu = (\mu_h)_{h \in [H]}$ such that for all $x_h \in \mathcal{X}_h$, $\mu_h(.|x_h)$ is an element of $\Delta(\mathcal{A}(x_h))$. A policy ν of the min-player can be defined similarly. We let $\Pi_{\max}$ and $\Pi_{\min}$ be the sets of the max- and min-player's policies.

Episode unfolding Given the two players' policies μ and ν, an episode of the game proceeds as follows: an initial state $s_1 \sim p_0$ is sampled. At step h, the max- and min-player observe their information sets x_h and y_h. Given the information, the max- and min-player choose and execute actions $a_h \sim \mu_h(\cdot|x_h)$ and $b_h \sim \nu_h(\cdot|y_h)$. As a result, the current state transitions to a next state $s_{h+1} \sim p_h(\cdot|s_h, a_h, b_h)$, and the max- and min-player receive rewards $r_h(s_h, a_h, b_h)$ and $-r_h(s_h, a_h, b_h)$ respectively. This is repeated until time step H, after which the episode finishes.

Learning procedure We assume there are T episodes of the same game, and both players are able to progressively adapt their respective policies μ^t and ν^t after each episode based on their previous observations.[4]

Realization plan and loss Given the perfect recall assumption, we recursively define (von Stengel, 1996) the realization plan $\mu_{1:} := (\mu_{1:h})_{h \in [H]}$ by, given $(x_1, a_1, ..., x_h)$ the unique history up to x_h,

$$\mu_{1:h}(x_h, a_h) := \prod_{h'=1}^{h} \mu_{h'}(a_{h'}|x_{h'}).$$

We also define the adversarial transitions $p_{1:}^{\nu} := (p_{1:h}^{\nu})_{h \in [H]}$ and losses $\ell^{\nu} := (\ell_h^{\nu})_{h \in [H]}$ by:

$$p_{1:h}^{\nu}(x_h) := p_0(x_1) \prod_{h'=1}^{h-1} p_{h'}^{\nu}(x_{h'+1}|x_{h'}, a_{h'}),$$
$$\ell_h^{\nu}(x_h, a_h) := p_{1:h}^{\nu}(x_h)(1 - r_h^{\nu}(x_h, a_h)),$$

where $p_h^{\nu}(x_{h+1}|x_h, a_h)$ is the probability to transition to the information state x_{h+1} from the pair information set-action (x_h, a_h) when the opponent policy is fixed to ν, and $r_h^{\nu}(x_h, a_h)$ is the average reward obtained in this case. Note that these two quantities require the perfect recall assumption in order to be well defined, and combine the randomness of both state transitions and min-player's policy ν.

---

[3]In many games with tree structure, p_h does not depend on the step h, but we keep the dependence on h as it makes the results more general at no extra cost in the analysis.

[4]As in Kozuno et al. (2021); Bai et al. (2022), we assume that the players can not change their policies within an episode, as the tree-like structure limits the interest of such adaptation.

The analog quantities $\nu_{1:}$, $p_{1:}^\mu$, r_h^μ and ℓ^μ may also be defined for the min-player, replacing $1 - r_h^\mu$ by r_h^μ in the definition of the loss.

Regret and ε-Nash-equilibrium We define the expected rewards (of the max-player) $V^{\mu,\nu} := \mathbb{E}^{\mu,\nu}\left[\sum_{h=1}^H r_h\right]$ for a pair $(\mu, \nu) \in \Pi_{\max} \times \Pi_{\min}$ of policies and, after T episodes, the regrets of the max and min players,

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

$$\mathfrak{R}_{\min}^T := \max_{\nu^\dagger \in \Pi_{\min}} \sum_{t=1}^T (V^{\mu^t, \nu^t} - V^{\mu^t, \nu^\dagger}).$$

There is a very clear link between minimizing the regret of both players and computing ε-optimal profiles, or ε-Nash-equilibrium (NE). In the particular case of perfect-recall, Theorem 1 by Kozuno et al. (2021) states the following (see also Cesa-Bianchi & Lugosi 2006; Zinkevich et al. 2007).

Theorem 2.1. *From $(\mu^t, \nu^t)_{t \in [T]}$ define the time-averaged profile $(\overline{\mu}, \overline{\nu})$ (of the realization plan, see Appendix E), then $(\overline{\mu}, \overline{\nu})$ is ε-optimal with*

$$\varepsilon = \left(\mathfrak{R}_{\max}^T + \mathfrak{R}_{\min}^T\right)/T.$$

Conversion to online linear regret minimization Defining a scalar product between the realization plans and the losses,

$$\langle \mu_{1:}, \ell^\nu \rangle := \sum_{h=1}^H \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \mu_{1:h}(x_h, a_h) \ell_h^\nu(x_h, a_h),$$

the chain rule on conditional probabilities implies that the max-player regret can then be rewritten, with $\ell^t := \ell^{\nu^t}$, as

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

This converts the problem to a linear regret problem, where the constraints set (the set of realization plans) is a polytope.

For simplicity, in the following sections we focus on the max-player although the exact same ideas apply to the min-player, who aims at minimizing $\mathfrak{R}_{\min}^T$.

3. Lower bounds

In this section we state lower bounds for learning an approximate NE. Compared to Bai et al. (2022), our lower bounds cover high-probability guarantees instead of in expectation, and are stated for two settings, whether or not the size of the action set is fixed. As shown in Section 4, they are tight up to logarithmic factors, which justifies the existence of two different rates for these settings.

Theorem 3.1. *(informal, exact statement in Appendix C)*

Regret: For any algorithm that controls the max-player, there exists a game such that, if the number of action per information set is allowed to vary, with probability at least δ, the algorithm suffers a regret of

$$\Omega\left(\sqrt{HA_\mathcal{X}T\log(1/(4\delta))}\right).$$

Otherwise if the number of the actions is fixed, then the regret is

$$\Omega\left(\sqrt{A_\mathcal{X}T\log(1/(4\delta))}\right).$$

Sample complexity: For any algorithm that controls both players, there exists a game such that, if the number of action is allowed to vary, it needs at least

$$\Omega\left(H(A_\mathcal{X} + B_\mathcal{Y})\log(1/(4\delta))/\varepsilon^2\right)$$

episodes to output an ε-optimal strategy with a probability at least $1 - \delta$.

If the number of actions is fixed, then it instead needs

$$\Omega\left((A_\mathcal{X} + B_\mathcal{Y})\log(1/(4\delta))/\varepsilon^2\right) \text{ episodes.}$$

These lower bounds are obtained by a reduction to the K-armed bandits. The structure depends on the setting: when the number of actions is allowed to vary, the learner must only choose between K actions on the first information set, and then gets H times the same reward depending on this first choice. When the number of actions is fixed, the learner must take H actions successively (among the A proposed at each information set) and only receive a reward after the last one depending on all of its choices; hence with $K = A^H$ possibilities.

4. Optimal rate with the knowledge of the tree structure

We first study the case where the information set structure is known by the agent beforehand.

Loss estimation The agent only observes the realizations of the games. This suggests that the loss vector ℓ^t should be estimated using only the information on the visited states and the associated rewards. Therefore, we define, for all possible actions, the unbiased loss estimate

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

where x_h^t a_h^t and r_h^t are respectively the information set, action and reward of the max-player at time step h of episode t.

4

While this estimator works well in expectation, its variance is too large to get acceptable bounds with high probability. We replace it by the following (biased) estimator $\widetilde{\ell}^t$, with an implicit exploration (IX) parameter γ_h that depends on the chosen state (Kocák et al., 2014):

$$\widetilde{\ell}_h^t(x_h, a_h) := \frac{\mathbb{I}_{\{x_h^t = x_h, a_h^t = a_h\}}}{\mu_{1:h}^t(x_h, a_h) + \gamma_h(x_h)}(1 - r_h^t).$$

The associated cumulative loss is given as $\widetilde{L}^t := \sum_{k=1}^{t} \widetilde{\ell}^k$.

Balanced transitions We follow the idea of Bai et al. (2022) and define a *balanced transition* kernel $p^\star$ on $\mathcal{X}$. First, we denote by $A_x^\tau = \sum_{x' \geq x} A_{x'}$ the total number of actions that follow information set $x \in \mathcal{X}$ in the whole sub-tree. The transition kernel $p^\star$ is then defined such as, at each step, the probability of transitioning to the next suitable information set x is proportional to this value. More precisely, $p_0^\star(x_1) = A_{x_1}^\tau / A_{\mathcal{X}}$ for all roots $x_1 \in \mathcal{X}_1$, and for any $h \in [H-1]$,

$$p_h^\star(x_{h+1}|x_h, a_h) := \frac{A_{x_{h+1}}^\tau}{\sum_{x'_{h+1} \in \mathcal{X}_{h+1}(x_h, a_h)} A_{x'_{h+1}}^\tau},$$

where $\mathcal{X}_{h+1}(x_h, a_h)$ denotes all the possible information sets of depth $h+1$ that follow (x_h, a_h).

This definition differs from the balanced transition $p^{\star h}$ in the appendix of Bai et al. (2022) for two reasons:

- It is defined globally for all depths h, using directly the whole tree instead of a truncated version up to each depth.
- It uses the reachable number of action instead of the reachable number of states, as the number of action at each set may not be constant.

Policy update (U1) For any transition kernel p^ν on $\mathcal{X}$ and max-player policy μ, the vector $p_{1:h}^\nu \cdot \mu_{1:h}$ defined by $p_{1:h}^\nu \cdot \mu_{1:h}(x_h, a_h) := p_{1:h}^\nu(x_h)\mu_{1:h}(x_h, a_h)$ is a probability distribution over the information set-action $\mathcal{A}(\mathcal{X}_h)$ at depth h. Along the lines of FTRL algorithms, we define the policy update of `Balanced FTRL` as

$$\mu^t = \arg\min_{\mu \in \Pi_{\max}} \left\langle \mu_{1:}, \widetilde{L}^{t-1} \right\rangle + \sum_{h=1}^{H} \Psi_h \left(p_{1:h}^\star \cdot \mu_{1:h} \right) / \eta_h$$

$$(\text{U1})$$

where $(\Psi_h)_{h \in [H]}$ are regularizers of $\mathcal{A}(\mathcal{X}_h)$ and $(\eta_h)_{h \in [H]}$ is a sequence of learning rates. While the ideal choice for regret minimization would be the actual adversarial transitions p^{ν^t} (replacing our $p^\star$) these transitions are unknown by the player. The balanced transitions instead give acceptable bounds against any adversarial transition. `Balanced FTRL` implements this idea, using either Tsallis entropy (Tsallis, 1988) or Shannon entropy as regularizer. Similarly, the IX parameters γ_h are also adjusted for each state with $\gamma_h^\star(x_h) = \gamma_h / p_{1:h}^\star(x_h)$.

Algorithm 1 `Balanced FTRL`-Tsallis/Shannon
1: **Input:**
     Tree-like structure of $\mathcal{A}(\mathcal{X})$
     Sequences $(\eta_h)_{h \in [H]}$ of learning rates
     Sequences $(\gamma_h)_{h \in [H]}$ of IX parameters
     Regularizers for $w_h \in \Delta(\mathcal{A}(\mathcal{X}_h))$:
     Tsallis: $\Psi_h(w_h) = - \sum_{(x_h, a_h)} w_h(x_h, a_h)^q$
     Shannon: $\Psi_h(w_h) = \sum_{(x_h, a_h)} w_h(x_h, a_h) \log(w_h(x_h, a_h))$
2: **Initialize:**
     For all depth h and $x_h \in \mathcal{X}_h$:
        $\gamma_h^\star(x_h) \leftarrow \gamma_h / p_{1:h}^\star(x_h)$
3: **For** $t = 1$ to T
     Compute update (U1):
$$\mu^t \leftarrow \arg\min_{\mu \in \Pi_{\max}} \left\langle \mu_{1:}, \widetilde{L}^{t-1} \right\rangle + \sum_{h=1}^{H} \Psi_h \left(p_{1:h}^\star \cdot \mu_{1:h} \right) / \eta_h$$
     For $h = 1$ to H
        Observe information set x_h^t
        Execute $a_h^t \sim \mu_h^t(.|x_h^t)$ and **receive** reward r_h^t
        Compute loss:
        $\widetilde{\ell}_h^t(x_h^t, a_h^t) \leftarrow (1 - r_h^t) / \left(\mu_{1:h}^t(x_h^t, a_h^t) + \gamma_h^\star(x_h^t) \right)$

Time complexity `Balanced FTRL` first needs to initially compute the balanced transitions, which can be done recursively from the bottom of the information set tree with a time complexity of $\mathcal{O}(A_{\mathcal{X}})$.

Then, with `Balanced FTRL-Tsallis`, each update (U1) can be done by solving a convex programming over $\mathbb{R}^{A_{\mathcal{X}}}$. While this implies a computation with a polynomial cost in $A_{\mathcal{X}}$, it is still quite expensive when $A_{\mathcal{X}}$ gets large. On the other hand, we will see in Section 5 that the update (U1) of `Balanced FTRL-Shannon` has an equivalent formulation with the dilated Shannon entropy. Using Algorithm 3 of Appendix F, each update can then be computed with a time complexity of $\mathcal{O}(HA)$, where A is an upperbound on the local number of actions any information set can have.

Theoretical analysis We now state the main theoretical results for `Balanced FTRL` with proof in Appendix E.

Theorem 4.1. *Let* $\delta \in (0, 1)$ *and* $\iota = \log(3A_{\mathcal{X}}/\delta)$. *Fix the IX parameters to* $\gamma_h = \sqrt{H\iota/(2A_{\mathcal{X}}T)}$ *for all* $h \in [H]$.

Using `Balanced FTRL-Tsallis`, $q \in (0, 1)$ *and constant learning rates* $\eta_h = \sqrt{2q(1-q)/T}(H/A_{\mathcal{X}})^{q-1/2}$, *with probability at least* $1 - \delta$,

$$\mathfrak{R}_{\max}^T \leq \left(\sqrt{2/(q(1-q))} + 3\sqrt{2\iota} \right) \sqrt{HA_{\mathcal{X}}T}.$$

Using `Balanced FTRL-Shannon` *and constant learning rates* $\eta_h = \sqrt{2H\log(A_{\mathcal{X}})/(A_{\mathcal{X}}T)}$, *with probability at least* $1 - \delta$,

$$\mathfrak{R}_{\max}^T \leq \left(\sqrt{2\log(A_{\mathcal{X}})} + 3\sqrt{2\iota} \right) \sqrt{HA_{\mathcal{X}}T}.$$

Remark 4.2. The best theoretical rate is obtained with Tsallis entropy with $q = 1/2$, similarly to Abernethy et al. (2015). Even though Shannon entropy has an extra $\sqrt{\log(A_{\mathcal{X}})}$ factor in the (expected) first term of the bound as in the multi-armed bandits setting, both entropies give a rate $\mathcal{O}(\sqrt{H \log(A_{\mathcal{X}}/\delta) A_{\mathcal{X}} T})$ with high-probability. Indeed, the expected term is dominated in both cases by the ι term from the high-probability analysis.

Using Theorem 2.1, we conclude that the lower bound $\widetilde{\mathcal{O}}(H(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$ on the sample complexity is matched up to logarithmic factors when both players use `Balanced FTRL`. In particular, `Balanced FTRL` improves by a factor H over the rate of `Balanced OMD` by Bai et al. (2022), see Table 1.

Corollary 4.3. *For any $\varepsilon > 0$ and $\delta \in (0,1)$, if both player run either `Balanced FTRL`-Tsallis/-Shannon with the parameters specified in Theorem 4.1, $(\overline{\mu}, \overline{\nu})$ is an ε-optimal profile with probability at least $1 - \delta$ as long as*

$$T \geq \mathcal{O}\left(H\left(A_{\mathcal{X}} + B_{\mathcal{Y}}\right)\log(1/\delta + A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2\right).$$

The average profile can be computed along the update of the current profile, see Kozuno et al. (2021), at the price of a final update with a $\mathcal{O}(A_{\mathcal{X}} + B_{\mathcal{Y}})$ time complexity.

Fixed action set size When the size of the action set is fixed, the number of information sets having a given $x \in \mathcal{X}$ in its history will increase at least exponentially as the depth increases. This can be exploited by `Balanced FTRL` with suitable learning rates and IX parameters that both decrease exponentially with the depth h in order to remove the H dependency in the upper bounds, again nearly matching the lower bounds $\widetilde{\mathcal{O}}((A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$ given by Theorem 3.1 for this setting. The results, proved in Appendix E, are stated for Shannon entropy:

Theorem 4.4. *Assume that the size of the action set is constant equal to A and let $\delta \in (0,1)$, $\iota = \log(3A_{\mathcal{X}}/\delta)$. Then, `Balanced FTRL`-Shannon with $\gamma_h = \sqrt{A^{H-h}\iota/(2A_{\mathcal{X}}T)}$ and $\eta_h = \sqrt{2A^{H-h}\log(A_{\mathcal{X}})/(A_{\mathcal{X}}T)}$ yields*

$$\mathfrak{R}_{\max}^T \leq \left(5\sqrt{\log(A_{\mathcal{X}})} + 8\iota\right)\sqrt{A_{\mathcal{X}}T}.$$

As in Corollary 4.3, this allows to compute an ε-optimal profile as long as long as

$$T \geq \mathcal{O}\left((A_{\mathcal{X}} + B_{\mathcal{Y}})\log(1/\delta + A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2\right).$$

5. Near-optimal rate without the knowledge of the tree structure

In this section we drop the assumption on the knowledge of the tree structure of the information states.

Dilated entropy and policy update (U2) For a given vector η^t of learning rates on $\mathcal{X}$, we define the weighted dilated Shannon divergence (Kroer et al., 2015) between two policies $\mu^1, \mu^0 \in \Pi_{\max}$ by

$$\mathcal{D}_{\eta^t}(\mu^1, \mu^0) := \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{\mu_{1:h}^1(x_h, a_h)}{\eta_h^t(x_h)} \log \frac{\mu_h^1(a_h|x_h)}{\mu_h^0(a_h|x_h)}.$$

We are now interested in the updates

$$\mu^t = \mathrm{argmin}_{\mu \in \Pi_{\max}} \left\langle \mu_{1:}, \tilde{L}^{t-1} \right\rangle + \mathcal{D}_{\eta^t}(\mu, \mu^0) \quad \text{(U2)}$$

for $\mu^0 \in \Pi_{\max}$ a base policy fixed over all iterations. As shown by Algorithm 3 of Appendix F, these updates enjoy an efficient computation with a $\mathcal{O}(HA)$ time complexity if the vectors η^{t-1} of learning rates only change at the visited information sets $(x_1^{t-1}, \ldots, x_H^{t-1})$.

Connections with the balanced algorithm While there is no general equivalence between the Shannon entropy and the dilated Shannon entropy, the updates (U1) of `Balanced FTRL-Shannon` may be put in the form of updates (U2). Precisely, using the fact that $p^\star$ is a transition kernel over $\mathcal{X}$, the updates (U1) of `Balanced FTRL` can be rewritten for all $t \in [T]$ as (see Proposition F.2)

$$\mu^t = \mathrm{argmin}_{\mu \in \Pi_{\max}} \left\langle \mu, \tilde{L}^{t-1} \right\rangle + D_{\eta^\star}(\mu, \mu^\star),$$

for some $\mu^\star \in \Pi_{\max}$ and vector $\eta^\star$ of learning rates computed with a $\mathcal{O}(A_{\mathcal{X}})$ time complexity.

Adversarial transitions estimation The balanced adversarial transitions $p_{1:h}^\star(x_h)$ used in the above learning rates $\eta_h^\star(x_h)$ are impossible to compute if the tree structure is unknown at the start of the game. In analogy to the loss estimation, we could instead use the estimates of the actual adversarial transitions $p_h^{\nu^t}(x_h)$. For any $(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)$, an unbiased estimate of $p_h^{\nu^t}(x_h)$ is indeed given by $\widehat{p}_{1:h}^t(x_h, a_h) := \mathbb{I}_{\left\{x_h = x_h^t, a_h = a_h^t\right\}}/\mu_{1:h}^t(x_h, a_h)$.

Once more, we shall use IX estimators to get bounds with high probability and not just in expectation. However, `Balanced FTRL` used the IX parameter $\gamma_h^\star(x_h) = \gamma/p_{1:h}^\star(x_h)$ which also can not be computed for the same reason. Our idea is to again replace these transitions by the same estimated transitions we are currently trying to define. This leads to the following recursive definitions,

$$\gamma_h^t(x_h, a_h) := \gamma/\left(1 + \widetilde{P}_{1:h}^{t-1}(x_h, a_h)\right),$$

$$\widetilde{p}_h^t(x_h, a_h) := \frac{\mathbb{I}_{\left\{x_h = x_h^t, a_h = a_h^t\right\}}}{\mu_{1:h}^t(x_h, a_h) + \gamma_h^t(x_h, a_h)},$$

where $\widetilde{P}^t := \sum_{k=1}^{t} \widetilde{p}^k$ are the cumulative estimated transitions and $\gamma_h^t(x_h, a_h)$ the IX parameter that now also depends on the action a_h and on the time t.

Algorithm 2 `Adaptive FTRL`

1: **Input:**
   Learning rate η and IX base parameter γ
   Base policy μ^0 (uniform by default)
2: **For** $t = 1$ to T :
   For all depth h and $x_h \in \mathcal{X}_h$:
   $\eta_h^t(x_h) \leftarrow \min_{x'_{h'} \geq x_h} \eta/\big(1 + \widetilde{P}_{1:h'}^{t-1}(x'_{h'})\big)$
   For all $a_h \in \mathcal{A}(x_h)$:
   $\gamma_h^t(x_h, a_h) \leftarrow \gamma/\big(1 + \widetilde{P}_{1:h}^{t-1}(x_h, a_h)\big)$
   Compute update (U2):
   $\mu^t \leftarrow \operatorname{argmin}_{\mu \in \Pi_{\max}} \left\langle \mu_{1:}, \tilde{L}^{t-1} \right\rangle + \mathcal{D}_{\eta^t}(\mu, \mu^0)$
   For $h = 1$ to H:
   Observe information set x_h^t
   Execute $a_h^t \sim \mu_h^t(.|x_h^t)$ and **receive** reward r_h^t
   $\widetilde{\ell}_h^t(x_h^t, a_h^t) \leftarrow (1 - r_h^t)/(\mu_{1:h}^t(x_h^t, a_h^t) + \gamma_h^t(x_h^t, a_h^t))$
   $\widetilde{p}_{1:h}^t(x_h^t, a_h^t) \leftarrow 1/(\mu_{1:h}^t(x_h^t, a_h^t) + \gamma_h^t(x_h^t, a_h^t))$

Adaptive learning rates   To replace the learning rate $\eta_h^\star(x_h)$ associated to each information set $x_h \in \mathcal{X}_h$, we thus use the cumulative estimated transitions. As at each round t, for each action $a_h \in \mathcal{A}(x_h)$, the quantity $\widetilde{P}_{1:h}^t(x_h, a_h)$ is an estimator of $P_{1:h}^t h(x_h)$, we define the estimator $\widetilde{P}_{1:h}^t(x_h)$, as the average of these estimations

$$\widetilde{P}_{1:h}^t(x_h) := \sum_{a_h \in \mathcal{A}(x_h)} \widetilde{P}_{1:h}^t(x_h, a_h)/A_{x_h}.$$

This estimator is then used to define a learning rate $\eta_h^t(x_h)$, mimicking the action of $p_h^\star(x_h)$ in $\eta_h^\star(x_h)$,

$$\eta_h^t(x_h) := \min_{x'_{h'} \geq x_h} \eta/\big(1 + \widetilde{P}_{1:h'}^{t-1}(x'_{h'})\big).$$

For technical reasons, these learning rates have to be increasing along a trajectory, hence the $\min$ operator. Note that this definition at round t does not depend on whether the yet unexplored information states are considered in this $\min$ operator, as these verify $\widetilde{P}_{1:h'}^{t-1}(x'_{h'}) = 0$.

Adaptive algorithm   `Adaptive FTRL` is based on update (U2) with the above vectors $(\eta^t)_{t \in [T]}$ of learning rates and IX estimators, which allow it is naturally adaptive to the initially unknown tree structure; see Algorithm 2 for a detailed description of `Adaptive FTRL`.

Time complexity   Unlike the `Balanced FTRL` algorithm, there is no initialization cost. Algorithm 3, in Appendix F, gives an efficient way of computing each update (U2) with a $\mathcal{O}(HA)$ time complexity. Moreover, as $\widetilde{P}$ only increases on visited information states, the vectors η^t and γ^t can be updated in-place with the same $\mathcal{O}(HA)$ time complexity.

Theoretical results   We show high-probability sample complexity bounds for `Adaptive FTRL`, but with an ad-

ditional $H \log(T) \log(1/\delta + T)$ factor on the sample complexity compared to the optimal rate. This result is proved in Appendix G.

Theorem 5.1. *Let $\iota' = \log(3A_{\mathcal{X}}T/\delta)$ and $\upsilon = 1 + \log_2(1 + T)$. For any $\delta \in (0, 1)$, using* `Adaptive FTRL` *with $\eta = 2\sqrt{\iota'T/(\upsilon A_{\mathcal{X}})}$, $\gamma = \sqrt{2\iota' HT/(\upsilon A_{\mathcal{X}})}$ and μ^0 the uniform policy yields with a probability at least $1 - \delta$,*

$$\mathfrak{R}_{\max}^T \leq 6H\sqrt{\iota'\upsilon A_{\mathcal{X}}T}.$$

As in Corollary 4.3, this leads to the computation of an ε-optimal profile as long as

$$T \geq \mathcal{O}\left(H^2(A_{\mathcal{X}} + B_{\mathcal{Y}})\log(1/\delta + T)\log(T)/\varepsilon^2\right).$$

The rate of `Adaptive FTRL` improves the one of IXOMD—up till now the algorithm with the best rate without the knowledge of the structure—by a factor at least $\min(X, Y)$.

Parameter tuning   `Adaptive FTRL` needs the knowledge of the $A_{\mathcal{X}}$ constant but only in order to tune the learning rate and the IX parameter. Fortunately, the doubling trick can be applied to $A_{\mathcal{X}}$ similarly to how it would be applied to T, which means that, up to an additional constant factor, this knowledge is not required for the theoretical guarantees.

6. Experiments

In this section we provide preliminary experiments for `Balanced FTRL` and `Adaptive FTRL` on simple games. The code used to generate these experiments is available at https://github.com/anon17893/IIG-tree-adaptation.

Games   We compare the algorithms on the following three standard benchmark tabular games: Kuhn poker (Kuhn, 1950), Leduc poker (Southey et al., 2005) and liars dice. We use the implementation of these games available in the OpenSpiel library (Lanctot et al., 2019).

Implementation details   We implemented IXOMD (Kozuno et al., 2021), `Balanced OMD` (Bai et al., 2022), `Balanced FTRL-Shannon` and `Adaptive FTRL`. We also implemented `Tweaked Adaptive FTRL`, a slight modification of `Adaptive FTRL`, in which the learning rates and IX parameters decrease depending on $(1 + \widetilde{P}^t)^{-1/2}$ instead of $\sqrt{T}/(1 + \widetilde{P}^t)$. This version gets better practical performances, at the price of not having tight theoretical guarantees.

These algorithms strongly rely on the learning rates for their performances. However, the learning rates obtained with the theoretical analysis of the above algorithms are only adjusted with respect to the worst possible 2-player games
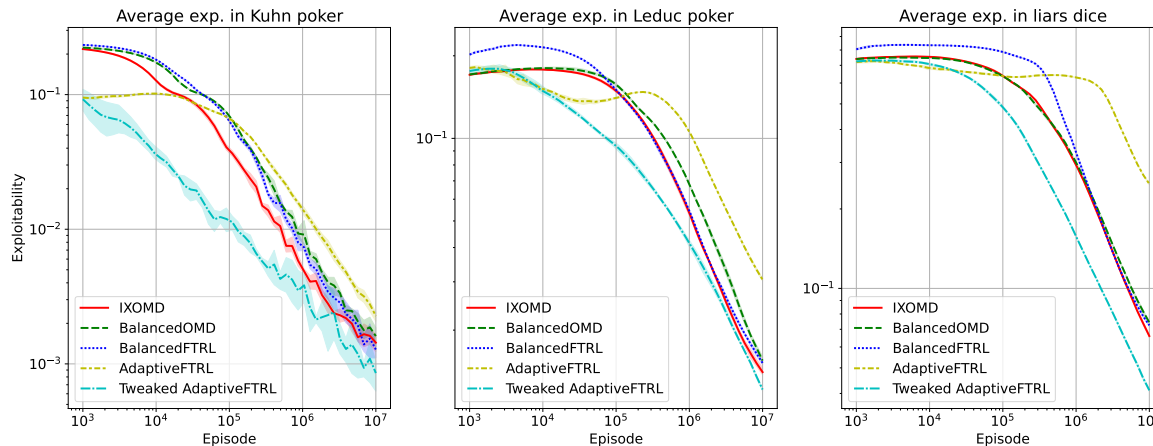

Figure 1. Performances of various algorithms with respect to the number of episodes for the time-averaged profile of Theorem 2.1. The results are given in terms of the exploitability gap $\max_{\mu \in \Pi_{\max}} V^{\mu, \overline{\nu}} - \min_{\nu \in \Pi_{\min}} V^{\overline{\mu}, \nu}$, with the rewards scaled between 0 and 1. The total number of actions is always the same for both players, with $A_{\mathcal{X}} = 12$ for Kuhn poker, $A_{\mathcal{X}} = 1092$ for Leduc poker, and $A_{\mathcal{X}} = 24570$ for Liars dice.

of a given size. With such rates, the practical performances are actually barely improved compared to the theoretical bounds. We thus tune the rates separately for each algorithm and each game, using a (logarithmic) grid search on the global learning rates, while the base IX parameter was taken as $1/20$ of this global learning rate.

Results Figure 1 shows the results of these experiments. We notice that, despite having very different theoretical guarantees, all algorithms seem to have very similar practical performances once tuned, except for `Adaptive FTRL` being a little worse and its modified version `Tweaked Adaptive FTRL` being slightly better on average. We believe that the behaviour of `Adaptive FTRL` can be explained by its learning rates of order $\sqrt{T}/(1 + \widetilde{P}^t)$, which is very large at the beginning leading to a fast convergence in the early phase. But the learning rates then need some time to decrease enough, as $\widetilde{P}^t$ has to grow, and allow `Adaptive FTRL` to make more progress [5]. `Tweaked Adaptive FTRL` stays adaptive but avoids this issue with a less steep learning rate decrease and a smaller initial learning rate.

7. Conclusion

We presented two algorithms. `Balanced FTRL` is problem-independent optimal for learning an ε-optimal profile. It requires to know the structure of the information set spaces in advance in order to compute a balanced transition kernel used to weight the regularizer. `Adaptive FTRL` does not require the knowledge of the information set space structure while being near problem-independent optimal. Indeed,

`Adaptive FTRL` directly estimates the transition kernels induced by the opponent instead of relying on a balanced transition kernel.

Our results bring the following interesting future directions:

Optimal rate for `Adaptive FTRL` Is it possible to save the extra H dependence over the horizon in the sample complexity of `Adaptive FTRL`?

Last iterate convergence Can we obtain theoretical guarantees for the current final profile instead of the average profile used in Theorem 2.1, in the same way as Lee et al. (2021)? Indeed, the computation of the time-averaged profile is not straightforward outside of the current tabular setting (Heinrich et al., 2015; Brown et al., 2019).

Avoiding importance sampling In large games, the importance exploration estimates tend to have a very high variance. This may be an issue if we want to combine the algorithms with function approximation (McAleer et al., 2022). Therefore, an interesting future direction would be to design algorithms that *do not rely on neither the importance sampling nor the knowledge of the information set structure, but still get sharp sample complexity guarantees*.

---

[5]Note that this effect is responsible for an additional $\sqrt{\log(T)}$ factor in the regret analysis

References

Abernethy, J. D., Lee, C., and Tewari, A. Fighting bandits with a new kind of smoothness. In *Neural Information Processing Systems*, 2015. URL https://proceedings.neurips.cc/paper/2015/file/5caf41d62364d5b41a893adc1a9dd5d4-Paper.pdf.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2159–2170. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/172ef5a94b4dd0aa120c6878fc29f70c-Paper.pdf.

Bai, Y., Jin, C., Mei, S., and Yu, T. Near-optimal learning of extensive-form games with imperfect information. In *International Conference on Machine Learning*, 2022.

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., Rowe, D., Shi, W., Spisak, J., Wei, A., Wu, D., Zhang, H., and Zijlstra, M. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/abs/10.1126/science.ade9097.

Brown, N. and Sandholm, T. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018. doi: 10.1126/science.aao1733. URL https://www.science.org/doi/abs/10.1126/science.aao1733.

Brown, N., Lerer, A., Gross, S., and Sandholm, T. Deep counterfactual regret minimization. In *International Conference on Machine Learning*, 2019. URL https://proceedings.mlr.press/v97/brown19b.html.

Burch, N., Moravčík, M., and Schmid, M. Revisiting CFR+ and Alternating Updates. *Journal of Artificial Intelligence Research*, 64:429–443, 2019.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006. ISBN 978-0-521-84108-5. doi: 10.1017/CBO9780511546921.

Daskalakis, C., Foster, D. J., and Golowich, N. Independent policy gradient methods for competitive reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5527–5540. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/3b2acfe2e38102074656ed938abf4ac3-Paper.pdf.

Farina, G., Kroer, C., and Sandholm, T. Regret circuits: Composability of regret minimizers. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1863–1872. PMLR, 2019. URL http://proceedings.mlr.press/v97/farina19b.html.

Farina, G., Kroer, C., and Sandholm, T. Stochastic Regret Minimization in Extensive-Form Games. In *International Conference on Machine Learning*, 2020.

Farina, G., Kroer, C., and Sandholm, T. Faster Game Solving via Predictive Blackwell Approachability: Connecting Regret Matching and Mirror Descent. In *AAAI Conference on Artificial Intelligence*, 2021a. URL https://arxiv.org/abs/2007.14358.

Farina, G., Kroer, C., and Sandholm, T. Bandit Linear Optimization for Sequential Decision Making and Extensive-Form Games. In *AAAI Conference on Artificial Intelligence*, 2021b.

Gilpin, A., Peña, J., and Sandholm, T. First-order algorithm with $O(ln(1/\varepsilon))$ convergence for ε-equilibrium in two-person zero-sum games. *Mathematical Programming*, 133, 2012. URL https://www.cs.cmu.edu/{~}sandholm/restart.MathProg12.pdf.

Gordon, G. J. No-regret Algorithms for Online Convex Programs. In *Advances in Neural Information Processing Systems*, 2007.

Hart, S. and Mas-Colell, A. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Heinrich, J., Lanctot, M., and Silver, D. Fictitious self-play in extensive-form games. In *International conference on machine learning*, 2015.

Hoda, S., Gilpin, A., Peña, J., and Sandholm, T. Smoothing Techniques for Computing Nash Equilibria of Sequential Games. *Mathematics of Operations Research*, 2010. URL https://kilthub.cmu.edu/ndownloader/files/12101699.

Johanson, M., Bard, N., Lanctot, M., Gibson, R., and Bowling, M. Efficient nash equilibrium approximation through

monte carlo counterfactual regret minimization. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '12, pp. 837–846, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0981738125.

Kocák, T., Neu, G., Valko, M., and Munos, R. Efficient learning by implicit exploration in bandit problems with side observations. In *Neural Information Processing Systems*, 2014.

Koller, D. and Pfeffer, A. Generating and solving imperfect information games. In *International Joint Conference on Artificial Intelligence*, 1995.

Koller, D., Megiddo, N., and Von Stengel, B. Efficient Computation of Equilibria for Extensive Two-Person Games. *Games and Economic Behavior*, 14(2):247–259, 1996.

Kozuno, T., Ménard, P., Munos, R., and Valko, M. Learning in two-player zero-sum partially observable Markov games with perfect recall. In *Neural Information Processing Systems*, 2021.

Kroer, C., Waugh, K., Kilinç-Karzan, F., and Sandholm, T. Faster first-order methods for extensive-form game solving. In *Economics and Computation*, 2015. ISBN 978-1-4503-3410-5. doi: 10.1145/2764468.2764476.

Kroer, C., Farina, G., and Sandholm, T. Solving large sequential games with the excessive gap technique. In *Neural Information Processing Systems*, 2018.

Kroer, C., Waugh, K., Kılınç-Karzan, F., and Sandholm, T. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179(1):385–417, 2020.

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

Kuhn, H. W. Extensive Games and the Problem of Information. *Annals of Mathematics Studies*, 28:193–216, 1953.

Lanctot, M., Waugh, K., Zinkevich, M., and Bowling, M. Monte-Carlo sampling for regret minimization in extensive games. In *Neural Information Processing Systems*, 2009.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., De Vylder, B., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., and Ryan-Davis, J. Openspiel:

A framework for reinforcement learning in games, 2019. URL https://arxiv.org/abs/1908.09453.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Lee, C.-W., Kroer, C., and Luo, H. Last-iterate convergence in extensive-form games. In *Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/file/77bb14f6132ea06dea456584b7d5581e-Paper.pdf.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 1994.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7001–7010. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21z.html.

McAleer, S., Farina, G., Lanctot, M., and Sandholm, T. ESCHER: Eschewing importance sampling in games by computing a history value function to estimate regret. *CoRR*, abs/2206.04122, 2022. doi: 10.48550/arXiv.2206.04122. URL https://doi.org/10.48550/arXiv.2206.04122.

Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960. URL https://www.science.org/doi/abs/10.1126/science.aam6960.

Munos, R., Pérolat, J., Lespiau, J.-B., Rowland, M., De Vylder, B., Lanctot, M., Timbers, F., Hennes, D., Omidshafiei, S., Gruslys, A., Azar, M. G., Lockhart, E., and Tuyls, K. Fast computation of nash equilibria in imperfect information games. In *International Conference on Machine Learning*, 2020.

Nemirovski, A. Prox-Method with Rate of Convergence $O(1/t)$ for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Neural Information Processing Systems*, 2015.

Pérolat, J., de Vylder, B., Hennes, D., Tarassov, E., Strub, F., de Boer, V., Muller, P., Connor, J. T., Burch, N., Anthony, T., McAleer, S., Elie, R., Cen, S. H., Wang, Z., Gruslys, A., Malysheva, A., Khan, M., Ozair, S., Timbers, F., Pohlen, T., Eccles, T., Rowland, M., Lanctot, M., Lespiau, J.-B., Piot, B., Omidshafiei, S., Lockhart, E., Sifre, L., Beauguerlange, N., Munos, R., Silver, D., Singh, S., Hassabis, D., and Tuyls, K. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 2022. doi: 10.48550/ARXIV.2206.15378. URL https://arxiv.org/abs/2206.15378.

Romanovsky, J. V. Reduction of a game with complete memory to a matricial game. *Dokl. Akad. Nauk SSSR*, 144:62–64, 1962.

Schmid, M., Burch, N., Lanctot, M., Moravčík, M., Kadlec, R., and Bowling, M. Variance reduction in monte carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. *CoRR*, abs/1809.03057, 2018. URL http://arxiv.org/abs/1809.03057.

Schmid, M., Moravčík, M., Burch, N., Kadlec, R., Davidson, J., Waugh, K., Bard, N., Timbers, F., Lanctot, M., Holland, Z., Davoodi, E., Christianson, A., and Bowling, M. Player of games. *arXiv preprint arXiv:2112.03178*, 2021.

Sidford, A., Wang, M., Yang, L., and Ye, Y. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2992–3002. PMLR, 2020. URL http://proceedings.mlr.press/v108/sidford20a.html.

Southey, F., Bowling, M., Larson, B., Piccione, C., Burch, N., Billings, D., and Rayner, C. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertaintyin Artificial Intelligence (UAI)*, pp. 550–558, 2005.

Strens, M. A Bayesian Framework for Reinforcement Learning. In *International Conference on Machine Learning*, 2000.

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

Tsallis, C. Possible Generalization of Boltzmann-Gibbs Statistics. *J. Statist. Phys.*, 52:479–487, 1988. doi: 10.1007/BF01016429.

von Neumann, J. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. ISSN 0025-5831; 1432-1807/e.

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

Waugh, K. and Bagnell, J. A. A unified view of large-scale zero-sum equilibrium computation. *CoRR*, abs/1411.5007, 2014. URL http://arxiv.org/abs/1411.5007.

Wei, C., Lee, C., Zhang, M., and Luo, H. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In Belkin, M. and Kpotufe, S. (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4259–4299. PMLR, 2021. URL http://proceedings.mlr.press/v134/wei21a.html.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/36e729ec173b94133d8fa552e4029f8b-Paper.pdf.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3674–3682. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/xie20a.html.

Zhang, B. H. and Sandholm, T. Finding and Certifying (Near-) Optimal Strategies in Black-Box Extensive-Form Games. In *AAAI Conference on Artificial Intelligence*, 2021.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1166–1178. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0cc6ee01c82fc49c28706e0918f57e2d-Paper.pdf.

Zhou, Y., Li, J., and Zhu, J. Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/pdf?id=Syg-ET4FPS.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. *Neural Information Processing Systems*, 2007.

Appendix

Table of Contents

A. Notation

Notation	Meaning
$\mathcal{S}$	state space of size S
p_h	state-transition probability kernel at step h
r_h	reward function at step h
H	length of one episode
$\mathcal{X}$	information set of the max-player of size X
$\mathcal{Y}$	information set of the min-player of size Y
$\mathcal{A}(x)$	action space at information set x of size A_x
$\mathcal{B}(y)$	action space at information set y of size B_y
$\mathcal{A}(\mathcal{X}) = \cup_h \mathcal{A}(\mathcal{X}_h)$	union of information set-action space $\mathcal{A}(\mathcal{X}_h)$
$\mathcal{B}(\mathcal{Y}) = \cup_h \mathcal{B}(\mathcal{Y}_h)$	union of information set-action space $\mathcal{B}(\mathcal{Y}_h)$
$\mathcal{X}_{h+1}(x_h, a_h)$	information sets of depth $h+1$ that follow (x_h, a_h)
$\mu_h(\cdot\|x_h)$	policy of the max-player at information-set x_h and step h
$\nu_h(\cdot\|y_h)$	policy of the min-player at information-set y_h and step h
$\mu_{1:h}(x_h, a_h) = \prod_{h'=1}^{h} \mu_{h'}(a_{h'}\|x_{h'})$	realization plan of the max-player
$p_h^\nu(x_{h+1}\|x_h, a_h)$	probability to transition to x_{h+1} from (x_h, a_h) with opponent ν
$r_h^\nu(x_h, a_h)$	average reward at (x_h, a_h) when the opponent is ν
$p_{1:h}^\nu(x_h) = p_0(x_1) \prod_{h'=1}^{h-1} p_{h'}^\nu(x_{h'+1}\|x_{h'}, a_{h'})$	adversarial realization plan against ν
$\ell_h^\nu(x_h, a_h) = p_{1:h}^\nu(x_h)\left(1 - r_h^\nu(x_h, a_h)\right)$	adversarial loss against ν
T	number of episodes
$\mathfrak{R}_{\max}^T$	regret of max-player
$\ell^t = \ell^{\nu^t}$	loss of the max-player at episode t
γ_h	IX parameter
$\widetilde{\ell}_h^t$	IX estimate of the loss of the max-player
$\widetilde{L}^t = \sum_{k=1}^{t} \widetilde{\ell}^k$	cumulative loss of the max-player
$p_{1:h}^\nu \cdot \mu_{1:h} = p_{1:h}^\nu(x_h)\mu_{1:h}(x_h, a_h)$	induced probability distribution over $\mathcal{A}(\mathcal{X}_h)$ at depth h
$p_h^\star$	balanced transition at step h
Ψ_h	regularizers on $\mathcal{A}(\mathcal{X}_h)$
$\gamma_h^\star(x_h) = \gamma_h / p_{1:h}^\star(x_h)$	adjusted IX parameter at information set x_h
$D_{\eta^t}(\mu_1, \mu_0)$	weighted (by η^t) dilated Shannon divergence on $\mathcal{X}$
$\gamma_h^t(x_h, a_h)$	adaptive IX parameter at (x_h, a_h)
$\widetilde{p}_h^t(x_h, a_h)$	estimated transition at (x_h, a_h)
$\widetilde{P}^t = \sum_{k=1}^{t} \widetilde{p}^k$	cumulative estimated transitions
$\eta_h^t(x_h)$	adaptive learning rate at x_h

Table 2. Table of notation use throughout the paper

B. Extended related work

In this section we review previous work on learning an ε-optimal strategies in IIGs.

Full feedback When the game is known, that is the information set structure space, transitions probability and reward function are provided a first line of work recasts the setting through the sequence-form representation of a game as a linear program which can be solved efficiently (Romanovsky, 1962; von Stengel, 1996; Koller et al., 1996). A second line of work relies on first-order optimization method for saddle point computation (Hoda et al., 2010; Kroer et al., 2015; 2018; 2020; Munos et al., 2020; Lee et al., 2021). In particular Hoda et al. (2010); Kroer et al. (2018) relies on the Nesterov smoothing technique Nesterov (2005) whereas Kroer et al. (2015; 2020) use the MirrorProx algorithm (Nemirovski, 2004). These methods has a rate of convergence of order $\widetilde{\mathcal{O}}(\text{poly}(H, A_\mathcal{X}, B_\mathcal{Y})/\varepsilon)$. Note that the rate is smaller than the lower bound of Theorem 3.1 since they assume full knowledge of the game. Even game dependent exponential rate can be obtained with these type of approach, see Gilpin et al. (2012) and Munos et al. (2020).

A third approach, counterfactual regret minimization (Zinkevich et al., 2007), leverages local regret minimisation, i.e. minimising a type of regret at each information set. Popular algorithm are based on the regret-matching algorithm (Hart & Mas-Colell, 2000; Gordon, 2007) such as CFR algorithm (Zinkevich et al., 2007) or based on a close variant of regret-matching, e.g. CFR+ (Tammelin, 2014; Burch et al., 2019; Farina et al., 2021a). Note that other local regret minimizers could be used, see for example Waugh & Bagnell (2014); Farina et al. (2019). These algorithms enjoy a guarantee of convergence of order $\widetilde{\mathcal{O}}(\text{poly}(H, A_{\mathcal{X}}, B_{\mathcal{Y}})/\varepsilon^2)$.

Nevertheless, all the methods described above need to *the information set tree* (or all the state space) or in order to compute one update. The cost of one traversal is of order $\mathcal{O}(X + Y)$ if the transitions and the actions of the other player are sampled; see for example the external-sampling MCCFR algorithm (Lanctot et al., 2009).

Trajectory feedback   A way to tackle the aforementioned issues is to consider the agnostic setting where the *agent has no prior knowledge of the game and only observes trajectories of the game*. Precisely, the rewards, the transition probabilities are unknown. This is the setting considered in this paper.

Model-based   A first method to deal with this limited feedback is to build a *model* of the game then run any full feedback algorithm in this model. For example, Zhou et al. (2020) use *posterior sampling* (PS, Strens, 2000) to learn a model and then use the CFR algorithm in games sampled from the posterior. They obtain a convergence rate of order $\widetilde{\mathcal{O}}(\text{poly}(H, S, A, B)/\varepsilon^2)$ but only when the games are actually sampled according to the known prior. Instead, Zhang & Sandholm (2021) rely on the principle of optimism in presence of uncertainty to incrementally build a model of the game. Then, the CFR algorithm is fed with *optimistic estimates* of the local regrets. They prove a high-probability sample complexity of order $\widetilde{\mathcal{O}}(\text{poly}(H, S, A, B)/\varepsilon^2)$.

Model-free   Another line of work (Lanctot et al., 2009; Johanson et al., 2012; Schmid et al., 2018; Farina et al., 2020) directly estimate the local regret via importance sampling that are then feed to the CFR algorithm. In particular, the outcome-sampling MCCFR (Lanctot et al., 2009; Farina et al., 2020) builds an importance sampling estimate of the counterfactual regret by playing according to a well-chosen *balanced policy*. Intuitively, this policy should ensure to *explore all the information sets*. Note that, depending on the structure of the information set space, playing uniformly over the actions at each information set is not necessarily a good choice. Instead, Farina et al. (2020) propose as a balanced policy to play action with probability proportional to the number of leaves in the sub-tree of possible next information sets. In particular, the outcome-sampling MCCFR algorithm requires the knowledge of the information set space structure to build its balanced policy. Nonetheless, in order to obtain ε-optimal strategies with high probability, MCCFR needs at most $\widetilde{\mathcal{O}}(H^3(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$ realizations of the game (Farina et al., 2020; Bai et al., 2022).

Latter, Kozuno et al. (2021) propose to combine *Online Mirror Descent (OMD)* with *dilated Shannon entropy as regularizer* and importance sampling estimate of the losses of a player, see also Farina et al. (2021b). They prove a sample complexity, for the proposed algorithm, IXOMD, of order $\widetilde{\mathcal{O}}(H^2(XA_{\mathcal{X}} + YB_{\mathcal{Y}})/\varepsilon^2)$. Interestingly, they do not need to know in advance the structure of the information set space to obtain this bound[6]. However, the sample complexity of IXOMD does not match the lower bound for this setting which is of order $\mathcal{O}((A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$, see Section 3 and Bai et al. (2022). Recently, Bai et al. (2022) propose the Balanced OMD algorithm that also relies on OMD but with a dilated entropy weighted by the realization plans of balanced policies as regularizers. These balanced policies generalize the one introduced by Farina et al. (2020) for all depths in the information set tree. For Balanced OMD, they prove a sample complexity of order $\widetilde{\mathcal{O}}(H^3(A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2)$ with an improved dependency on the sizes of the information set spaces at the price of a worse dependence in the horizon H.

Perfect information Markov game   Another line of work consider Markov game Kuhn (1953) with *perfect* information and limited feedback. However, they do not assume perfect recall. Sidford et al. (2020); Zhang et al. (2020); Daskalakis et al. (2020); Wei et al. (2021) consider the case where a *generative model* is available whereas Wei et al. (2017); Bai et al. (2020); Xie et al. (2020); Liu et al. (2021) deal with the *trajectory feedback* case. Although this setting is related to ours there is no direct comparison between the two.

---

[6]They only need to know the size of the information set spaces to tune optimally the learning rate

C. Lower bound proof details

To formally state lower bounds, we start with a tad of additional notation. For the proofs of the lower bounds, we consider games in which one player (say the min-player) has no effect on both rewards and state-transition dynamics. Therefore, we omit random variables related to the min-player.

An algorithm is a sequence $\mathcal{L} \coloneqq (\mathcal{L}^t)_{t=1}^{T+1}$, where $\mathcal{L}^t$ is a measurable mapping from $\mathcal{A}(\mathcal{X})^{(t-1)H} \times \mathbb{R}^{(t-1)H}$ to $\Pi_{\max}$. In t-th episode, the algorithm outputs a policy $\mu^t = \mathcal{L}^t(H_t)$ given the history $H_t \coloneqq (x_h^u, a_h^u, r_h^u)_{u=1,h=1}^{t-1,H}$ and follows it. By Ionescu–Tulcea theorem (Lattimore & Szepesvári, 2020, Theorem 3.3), there exists a probability space consistent with this procedure. After T-th episode, the algorithm outputs a final policy $\mu^{T+1} = \mathcal{L}^{T+1}(H_{T+1})$.

We prove the following theorems. Note that we assume deterministic rewards, but lower bounds show that the difficulty of IIGs is the same as the difficulty of bandit problems when $H = 1$. This is because we made use of the partial observability such that rewards look stochastic from the view point of players.

Theorem C.1 (Regret Lower Bound; Variable Action Space Size). *Fix horizon H, number of episodes T, total number of actions $A_{\mathcal{X}}$, and a positive scalar $\delta \in (0,1)$. If $A_{\mathcal{X}} < H$, there is no such game. If $A_{\mathcal{X}} \geq H$, $\delta < 1/4$, and $T \geq -0.4K \log(4\delta)$, for any algorithm, there exists a game such that with probability greater than δ, the algorithm suffers from the regret of*

$$\Omega \left(\sqrt{(A_{\mathcal{X}} - H) \min\{A_{\mathcal{X}} - H, H\} T \log\left(1/(4\delta)\right)} \right) .$$

Theorem C.2 (Sample Complexity Lower Bound; Variable Action Space Size). *Fix horizon H, total numbers of actions $A_{\mathcal{X}}$, $B_{\mathcal{Y}}$, and positive scalars $\delta \in (0,1)$, $\varepsilon \in (0, H)$. If either $A_{\mathcal{X}} < H$ or $B_{\mathcal{Y}} < H$, there is no such game. If $A_{\mathcal{X}} \geq H$, $B_{\mathcal{Y}} \geq H$, and $\delta < 1/4$, for any algorithm, there exists a game such that the algorithm needs at least*

$$\Omega \left(\frac{(A_{\mathcal{X}} - H) \min\{A_{\mathcal{X}} - H, H\} + (B_{\mathcal{Y}} - H) \min\{B_{\mathcal{Y}} - H, H\}}{\varepsilon^2} \log\left(1/(4\delta)\right) \right)$$

episodes to output ε-NE with probability at least $1 - \delta$.

Theorem C.3 (Regret Lower Bound; Fixed Action Space Size). *Fix horizon H, number of episodes T, number of actions A at each information set, number of information sets X, and a positive scalar $\delta \in (0,1)$. If $X < (A^H - 1)/(A-1)$, there is no such game. If $X < 2(A^H - 1)/(A-1)$, $\delta < 1/4$, and $T \geq -0.4K \log(4\delta)$, for any algorithm, there exists a game such that with probability greater than δ, the algorithm suffers from the regret of*

$$\Omega \left(\sqrt{T A_{\mathcal{X}} \log\left(1/(4\delta)\right)} \right) .$$

Theorem C.4 (Sample Complexity Lower Bound; Fixed Action Space Size). *Fix horizon H, number of actions A, B at each information set, numbers of information sets X, Y, and positive scalars $\delta \in (0,1)$, $\varepsilon \in (0, H)$. If either $X < (A^H - 1)/(A-1)$ or $B < (B^H - 1)/(B-1)$, there is no such game. If $(A^H - 1)/(A-1) \leq X < 2(A^H - 1)/(A-1)$, $(B^H - 1)/(B-1) \leq Y < (B^H - 1)/(B-1)$, and $\delta < 1/4$, for any algorithm, there exists a game such that the algorithm needs at least*

$$\Omega \left(\frac{A_{\mathcal{X}} + B_{\mathcal{Y}}}{\varepsilon^2} \log\left(1/(4\delta)\right) \right)$$

episodes to output ε-NE with probability at least $1 - \delta$.

C.1. Variable action space sizes

We first consider a case in which the number of viable actions depends on information sets.

Proof sketch   In the hard game instance we consider, we make actions of one player ineffective depending on which one of $A_{\mathcal{X}}$ and $B_{\mathcal{Y}}$ is bigger. Specifically, if $A_{\mathcal{X}} > B_{\mathcal{Y}}$, the min-player's actions have no effect on the state-transition dynamics and rewards.[7] Then the game reduces to a Markov decision process, and finding an ε-optimal policy is equivalent to finding an ε-NE. For the time being, we assume that $A_{\mathcal{X}} > B_{\mathcal{Y}}$.

---

[7]In case of ties, make actions of the min-player ineffective.

Due to the perfect-recall assumption (at least one successive information set for each action except the final time step), $A_\mathcal{X}$ must be greater than or equal to H. Furthermore, when $A_\mathcal{X} = H$, there is only one policy, and the sample complexity lower bound is trivially $0 = A_\mathcal{X} - H$. Hereafter we assume that $A_\mathcal{X} > H$.
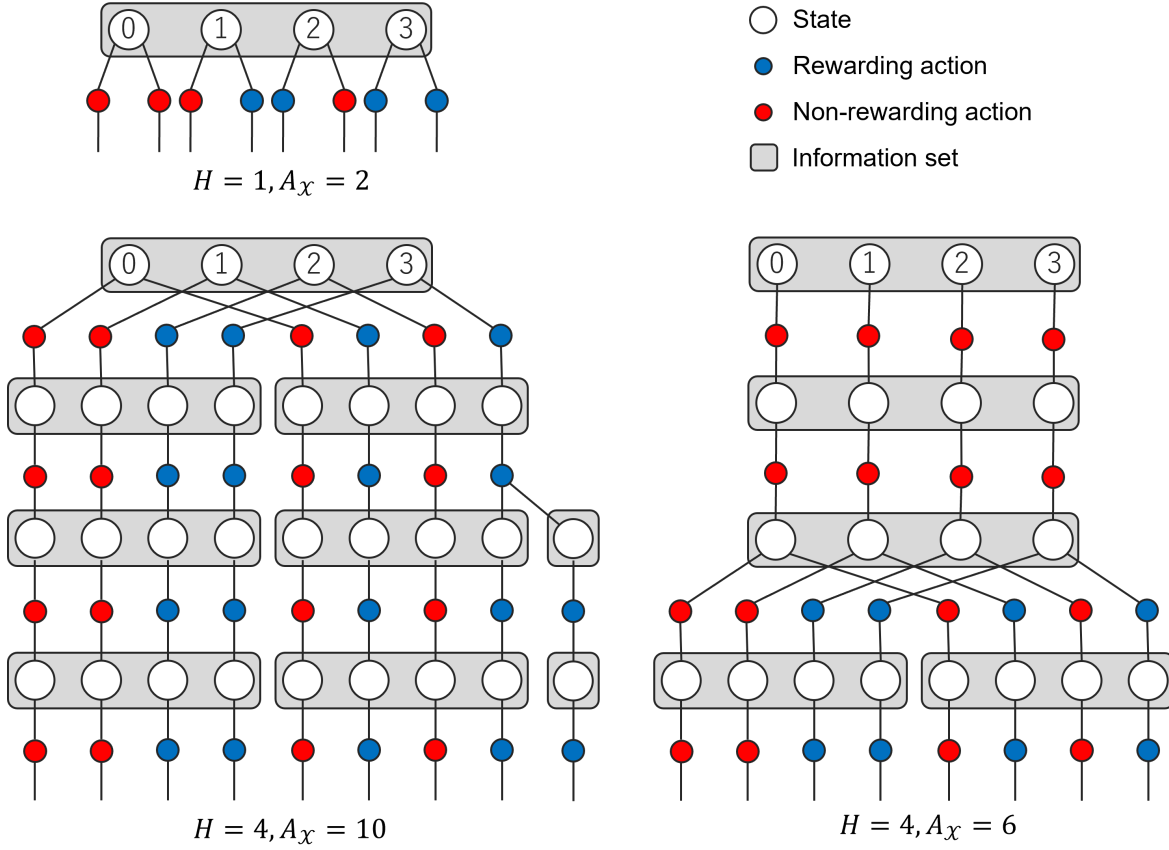

Figure 2. A hard game instance with information-set-dependent action space from the view point of the max-player.

Figure 2 depicts hard game instances under different conditions. White circles are states, and states at the top (and identified by integers) are initial states. Edges from each state are viable actions at a corresponding state. Blue circles are rewarding actions, where the max-player receives the reward of 1. Red circles are non-rewarding actions. Shaded boxes are information sets, and all states in a box belong to the same information set.

When $A_\mathcal{X} \geq 2H$, we construct a game shown in the bottom left panel. For brevity let $K := \lfloor A_\mathcal{X}/H \rfloor \geq A_\mathcal{X}/H - 1$ and assume that $A_\mathcal{X}$ is divisible by H. In this case there is no right-most branch. If $A_\mathcal{X}$ is not divisible by H, append a small branch as shown in the figure and set state-transition probability to the small branch 0. In this game, there are 2^K initial states, at which there are K viable actions. Each initial state is identified by an integer from 0 to $2^K - 1$. An initial state is sampled as follows: in the beginning of each episode, K binary integers are sampled from K (ordered) Bernoulli distributions and concatenated to construct a binary representation of an integer from 0 to $2^K - 1$ (e.g., 010111), which is an initial state in this episode. The parameter of k-th Bernoulli distribution is denoted by p_k. The reward value for k-th action is the k-th bit of the initial state in binary representation. After the first step, the state transitions in a deterministic way as shown in the figure, and reward values at later time steps are the same as the reward value at the first time step.

To get an intuition about this hard game instance, it is beneficial to consider a case where $H = 1$. In this case, the game essentially reduces to a bandit problem shown in the top left panel of Figure 2. In this case the regret lower bound of $\Omega(\sqrt{KT})$ and the sample complexity lower bound of $\Omega(K/\varepsilon^2)$ are known to hold. When the reward is scaled by H, the lower bounds become $\Omega(\sqrt{H(A_\mathcal{X} - H)T})$ and $\Omega(H(A_\mathcal{X} - H)/\varepsilon^2)$, respectively. The hard game instance described above instantiates this idea.

The hard game construction described above does not work when $H > 1$ and $2H > A_\mathcal{X} > H$. In this case, we construct a

game shown in the bottom right panel. The idea is almost the same, but there are only 4 initial states, and the player makes decision at $(2H - A_{\mathcal{X}} + 1)$-th time step instead of the first time step. There the max-player has only 2 actions, and reward values at $(2H - A_{\mathcal{X}} + 1)$-th time step and thereafter are determined as before. This game is essentially a 2-arm Bernoulli bandit with reward scaled by $A_{\mathcal{X}} - H$. In this case the regret lower bound of $\Omega((A_{\mathcal{X}} - H)\sqrt{2T})$ and the sample complexity lower bound of $\Omega(2(A_{\mathcal{X}} - H)^2 / \varepsilon^2)$ are known to hold.

If $A_{\mathcal{X}} > B_{\mathcal{Y}}$ does not hold, we can derive the same result with $A_{\mathcal{X}}$ replaced by $B_{\mathcal{Y}}$ by exchanging the role of max- and min-players in the previous argument. Consequently, depending on which one of $A_{\mathcal{X}}$ and $B_{\mathcal{Y}}$ is bigger, we can use $A_{\mathcal{X}}$ or $B_{\mathcal{Y}}$ in lower bounds. The claimed results follow from a fact that $\max\{a, b\} \geq (a + b)/2$.

Full Proof. We assume that $A_{\mathcal{X}} \geq 2H$ as other cases can be handled similarly. We consider $\lfloor A_{\mathcal{X}}/H \rfloor + 1$ games and identify each game by an integer from 0 to $\lfloor A_{\mathcal{X}}/H \rfloor$. In 0-th game , we set all p_k to 0.5. In k-th game, we set $(p_j)_{j \neq k}$ to 0.5, and p_k to $0.5 + \Delta$, where $\Delta \in (0, 0.5)$ is specified later in the proof. In other words, an optimal arm in k-th game is k-th action. We let $\mathbb{P}_k$ denote the probability measure induced by the interconnection of the algorithm and game k. Expectation under $\mathbb{P}_k$ is denoted by $\mathbb{E}_k$. We also let $\mathbb{P}'_k$ and $\mathbb{E}'_k$ denote the law of $(x_h^u, a_h^u, r_h^u)_{u=1, h=1}^{T,H}$ and expectation under $\mathbb{P}'_k$ when the algorithm is run in game k.

The proofs rely on the following core lemma.

Lemma C.5. *Fix scalars $\rho \in (0, 1)$ and $v \in (0, \infty)$. For brevity let $K := \lfloor A_{\mathcal{X}}/H \rfloor$. If $\Delta \leq 0.3$, for any measurable function $f : \mathcal{A}(\mathcal{X})^{TH} \times \{0, 1\}^{TH} \to \{0, 1\}$ and $k \in [K]$, it holds that*

$$\mathbb{P}_k\left(f(H_{T+1}) = 1\right) > \exp\left(-s - v\right)\left(\mathbb{P}_0\left(f(H_{T+1}) = 1\right) - \rho - \frac{4\Delta^2}{v} n_k\right),$$

where $s := \dfrac{4\Delta}{3}\log\dfrac{1}{\rho} + \sqrt{2v\log\dfrac{1}{\rho}}$ and $n_k := \mathbb{E}_0\left[\sum_{t=1}^T \mathbb{I}_{\{a_1^t = k\}}\right]$.

Proof. Let E be an event $\{f(H_{T+1}) = 1\}$. We begin with rewriting $\mathbb{P}_k(E)$ as follows:

$$\mathbb{P}_k(E) = \mathbb{E}_k\left[f(H_{T+1})\right] = \mathbb{E}'_k\left[f(H_{T+1})\right] = \mathbb{E}'_0\left[f(H_{T+1})\frac{d\mathbb{P}'_k}{d\mathbb{P}'_0}(H_{T+1})\right]. \tag{1}$$

We derive a lower bound for $d\mathbb{P}'_k/d\mathbb{P}'_0(H_{T+1})$. Because $\mathcal{A}(\mathcal{X})^{TH} \times \{0, 1\}^{TH}$ is discrete, $d\mathbb{P}'_k/d\mathbb{P}'_0(H_{T+1})$ is simply an importance sampling ratio. Furthermore, changing the game from 0 to k only changes the reward distribution when k-th action is taken. Therefore,

$$\frac{d\mathbb{P}'_k}{d\mathbb{P}'_0}(H_{T+1}) = \prod_{t=1}^T \left[r_1^t\left(\frac{0.5 + \mathbb{I}_{\{a_1^t = k\}}\Delta}{0.5}\right) + (1 - r_1^t)\left(\frac{0.5 - \mathbb{I}_{\{a_1^t = k\}}\Delta}{0.5}\right)\right]$$

$$= \prod_{t=1}^T \left[r_1^t\left(1 + 2\mathbb{I}_{\{a_1^t = k\}}\Delta\right) + (1 - r_1^t)\left(1 - 2\mathbb{I}_{\{a_1^t = k\}}\Delta\right)\right] = \prod_{t=1}^T (1 + P_t),$$

where $P_t := 2\Delta\left(2r_1^t - 1\right)\mathbb{I}_{\{a_1^t = k\}}$. From a fact that $1 + x \geq \exp\left(x - x^2\right)$ for $x \geq -0.6$, we deduce that

$$\frac{d\mathbb{P}'_k}{d\mathbb{P}'_0}(H_{T+1}) \geq \exp\left(-S_T - V_T\right),$$

where we defined $S_T := -\sum_{t=1}^T P_t$ and $V_T := \sum_{t=1}^T P_t^2 = 4\Delta^2 \sum_{t=1}^T \mathbb{I}_{\{a_1^t = k\}}$.

Returning to equation (1), it clearly holds that

$$\mathbb{P}_k(E) \geq \mathbb{E}'_0\left[f(H_{T+1})\mathbb{I}_{\{S_T < s, V_T \leq v\}}\frac{d\mathbb{P}'_k}{d\mathbb{P}'_0}(H_{T+1})\right]$$

$$> \exp\left(-s - v\right)\mathbb{E}'_0\left[f(H_{T+1})\mathbb{I}_{\{S_T < s, V_T \leq v\}}\right]$$

$$= \exp\left(-s - v\right)\mathbb{E}_0\left[f(H_{T+1})\mathbb{I}_{\{S_T < s, V_T \leq v\}}\right].$$

18

We lower-bound $f(H_{T+1})\mathbb{I}_{\{S_T<s,V_T\leq v\}}$. To this end, note that $A\cap B=(A\cup B^c)\cap B$ for any sets A and B. Therefore,

$$f(H_{T+1})\mathbb{I}_{\{S_T<s,V_T\leq v\}}=f(H_{T+1})\left(1-\mathbb{I}_{\{S_T\geq s,V_T\leq v\}}-\mathbb{I}_{\{V_T>v\}}\right)$$
$$\geq f(H_{T+1})-\mathbb{I}_{\{S_T\geq s,V_T\leq v\}}-\mathbb{I}_{\{V_T>v\}},$$

and thus,

$$\mathbb{E}_0\left[f(H_{T+1})\mathbb{I}_{\{S_T<s,V_T\leq v\}}\right]\geq\mathbb{P}_0\left(E\right)-\underbrace{\mathbb{P}_0\left(S_T\geq s,V_T\leq v\right)}_{\leq\rho\text{ by (a)}}-\underbrace{\mathbb{P}_0\left(V_T>v\right)}_{\leq\mathbb{E}_0[V_T]/v\text{ by (b)}}.$$

It is easy to see that (b) follows from Markov's inequality. To see why (a) follows, let $\mathcal{F}'_t$ be a σ-algebra generated by $(s^u_h,a^u_h)^{t-1,H}_{u=1,h=1}$ and a^t_1. Then, (a) follows from Freedman's inequality by noting that $(P_t)^n_{t=1}$ is a martingale difference sequence with respect to the filtration $(\mathcal{F}'_t)^T_{t=1}$, P_t is almost surely bounded by 2Δ, and $\sum^T_{t=1}\mathbb{E}_0[(P_t-\mathbb{E}_0[P_t|\mathcal{F}'_t])^2|\mathcal{F}_t]=V_T$ as $\mathbb{E}_0[P_t|\mathcal{F}'_t]=0$. Consequently,

$$\mathbb{P}_k(E)>\exp\left(-s-v\right)\left(\mathbb{P}_0\left(E\right)-\rho-\frac{4\Delta^2}{v}\mathbb{E}_0\left[\sum^T_{t=1}\mathbb{I}_{\{a^t_1=k\}}\right]\right)=\exp\left(-s-v\right)\left(\mathbb{P}_0\left(E\right)-\rho-\frac{4\Delta^2}{v}n_k\right).$$

This concludes the proof. $\qquad\square$

Regret Lower Bound. Now we are ready to prove the regret lower bound. For each $k\in[K]$, let E_k be an event $\sum^T_{t=1}\mu^t_1(a_k|x_1)\leq 0.5T$. Note that while implicit, $\sum^T_{t=1}\mu^t_1(a^*_k|x^t_1)$ is a function of H_{T+1}, and thus, E_k can be rewritten in a form of $f(H_{T+1})=1$. As we verify later, our choice of Δ is smaller than 0.3. Therefore from the previous lemma,

$$\max_{k\in[K]}\mathbb{P}_k(E_k)\geq\frac{1}{K}\sum_{k\in[K]}\mathbb{P}_k(E_k)>\exp\left(-s-v\right)\left(\frac{1}{K}\sum^K_{k=1}\mathbb{P}_0\left(E_k\right)-\rho-\frac{4T\Delta^2}{Kv}\right).$$

Note that $\sum_{k\in[K]}\mathbb{P}_0\left(E_k\right)=\mathbb{E}_0\left[\sum_{k\in[K]}\mathbb{I}_{\{E_k\}}\right]\geq K-1$ since $\mathbb{I}_{\{E_k\}}=0$ for some k implies that $\mathbb{I}_{\{E_j\}}=1$ for $j\neq k$. Thus,

$$\max_{k\in[K]}\mathbb{P}_k(E_k)>\exp\left(-s-v\right)\left(1-\frac{1}{K}-\rho-\frac{4T\Delta^2}{Kv}\right).$$

Setting $v=\dfrac{32T\Delta^2}{K}$ and $\rho=\dfrac{1}{8}$,

$$\max_{k\in[K]}\mathbb{P}_k(E_k)>\frac{1}{4}\exp\left(-4\Delta\log 2-8\Delta\sqrt{\frac{3T}{K}\log 2}-\frac{32T\Delta^2}{K}\right),$$

where we used an assumption that $K\geq 2$. Equating the right hand side to δ and solving for Δ, we choose

$$\Delta=\sqrt{\left(\frac{K}{16T}\log 2+\frac{1}{8}\sqrt{\frac{3K}{T}\log 2}\right)^2+\frac{K}{32T}\log\frac{1}{4\delta}}-\frac{K}{16T}\log 2-\frac{1}{8}\sqrt{\frac{3K}{T}\log 2}<\sqrt{\frac{K}{32T}\log\frac{1}{4\delta}}<0.3,\quad(2)$$

where the middle inequality follows since $\sqrt{a+b}<\sqrt{a}+\sqrt{b}$ for $a,b\in(0,\infty)$. Then it holds that $\max_{k\in[K]}\mathbb{P}_k(E_k)>\delta$. Because the event E_k implies that the regret is greater than or equal to $0.5T\Delta H$ in k-th game, we have a lower bound $\Omega(\sqrt{-H^2TK\log(4\delta)})=\Omega(\sqrt{-H^2T(A_{\mathcal{X}}/H-1)\log(4\delta)})=\Omega(\sqrt{-HT(A_{\mathcal{X}}-H)\log(4\delta)})$.

Sample Complexity Lower Bound. Now we are ready to prove the sample complexity lower bound. Let E_k be an event $\mu^{T+1}_1(a_k|x_1)\leq 0.5$ for each $k\in[K]$. By the same argument as the one in the regret lower bound proof, $\max_{k\in[K]}\mathbb{P}_k(E_k)>\delta$ for Δ in Equation 2. Since in k-th game, the event E_k implies that the simple regret of μ^{T+1} is greater than or equal to $0.5\Delta H$, we have a lower bound for the simple regret of $\Omega(\sqrt{-H(A_{\mathcal{X}}-H)/T\log(4\delta)})$. Therefore unless $T\geq-H(A_{\mathcal{X}}-H)/\varepsilon^2\log(4\delta)$, the algorithm fails to output ε-optimal policy with probability greater than δ.

C.2. Fixed action space size

Now we turn to the case where the action space depends on information sets. Since the proof is relatively easy, we directly give a full proof.

In the hard game instance we will consider, we make actions of one player ineffective depending on which one of $A_\mathcal{X}$ and $B_\mathcal{Y}$ is bigger, as before. For the time being, we assume that $A_\mathcal{X} > B_\mathcal{Y}$. Furthermore we restrict rewards to be binary but allow them to be stochastic. As we did in the proof of Lemma C.5, we can create a game in which rewards are deterministic but look stochastic from the view point of players.

First of all, we claim that there is no game such that $X < (A^H - 1)/(A - 1)$. This is because $X \geq 1 + A + \cdots + A^{H-1} = (A^H - 1)/(A - 1)$ due to the perfect-recall assumption (at least one successive information set for each action except the final time step). Hereafter we assume that $X \geq (A^H - 1)/(A - 1)$.
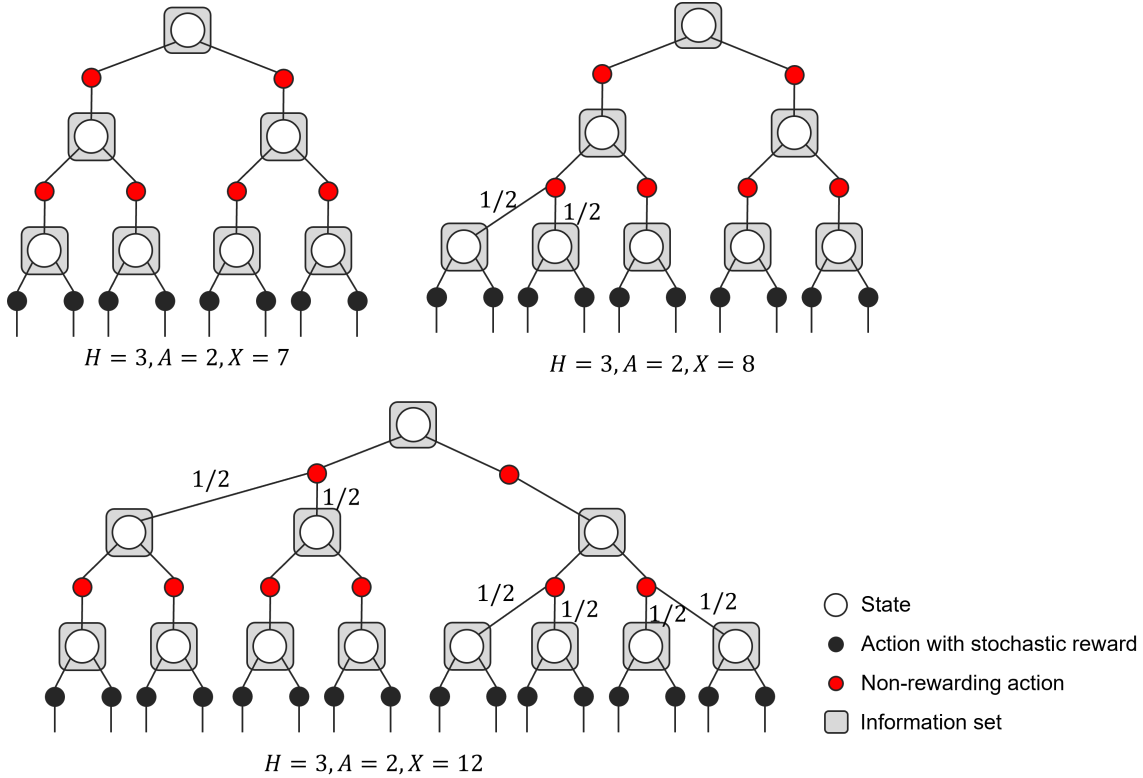

$H = 3, A = 2, X = 7$

$H = 3, A = 2, X = 8$

$H = 3, A = 2, X = 12$

○ State

● Action with stochastic reward

● Non-rewarding action

▢ Information set

Figure 3. A hard game instance with information-set-independent action space from the view point of the max-player.

Figure 3 depicts the hard game instance under different conditions. In all cases, all information sets are singleton. Unless a number is given, all edges from action nodes to state nodes mean deterministic transitions. Numbers along edges indicate state-transition probabilities of the edges. In all cases, the max-player receives reward 1 at leaf action nodes with some probability specified later in the proof.

The top left panel shows the case where $X = (A^H - 1)/(A - 1)$. In this case, the game tree is A-ary tree with deterministic transitions. This case is dealt by Bai et al. (2022), and we obtain a similar (pseudo-)regret bound.

The top right panel shows the case where $X = (A^H - 1)/(A - 1) + 1$. In this case, an additional A-ary tree with depth 1 is appended at the left-most action node at $(H - 1)$-th time step. State-transition probabilities to the successor state nodes are 0.5. Up to $(A^H - 1)/(A - 1) + A^{H-1}$, additional A-ary trees with depth 1 are added similarly to different action nodes at $(H - 1)$-th time step one by one from left to right action nodes. For example, if $X = (A^H - 1)/(A - 1) + 2$, another additional A-ary tree with depth 1 is appended at the second left-most action node at $(H - 1)$-th time step.

The bottom panel shows the case where $X = (A^H - 1)/(A - 1) + A^{H-1} + 1$. In this case, we remove the additional A-ary tree at the left-most action node at $(H - 1)$-th time step and append an A-ary tree with depth 2 at the left-most

action node at $(H-2)$-th time step. Up to $X = (A^H - 1)/(A-1) + A^{H-1} + A^{H-2}$, additional A-ary trees with depth 1 are added similarly to different action nodes at $(H-2)$-th time step one by one from left to right action nodes. This procedure can be repeated until all action nodes at the first time step have additional branches. At this point, there are $(A^H - 1)/(A-1) + A^{H-1} + \cdots + A = 2(A^H - 1)/(A-1) - 1$ information sets.

Now we specify reward probabilities at leaf action nodes and construct $A^H + A \min\{X - (A^H - 1)/(A-1), A^{H-1}\} + 1$ games. We index leaf action nodes by integers from 1 to $2A^H$. We let a_i^* and x_i^* denote the corresponding action of i-th node and its predecessor information set (i.e., an information set at which a_i^* can be taken). In 0-th game , we set all reward probabilities to 0.5. In k-th game, we set reward probabilities of all leaf action nodes except k-th one to 0.5, and the reward probability of k-th leaf action node to $0.5 + \Delta$, where $\Delta \in (0, 0.5)$ is specified later in the proof. In other words, an optimal leaf action node in k-th game is k-th one. We let $\mathbb{P}_k$ denote the probability measure induced by the interconnection of the algorithm and game k. Expectation under $\mathbb{P}_k$ is denoted by $\mathbb{E}_k$. We also let $\mathbb{P}'_k$ and $\mathbb{E}'_k$ denote the law of $(x_h^u, a_h^u, r_h^u)_{u=1, h=1}^{T, H}$ and expectation under $\mathbb{P}'_k$ when the algorithm is run in game k.

By using the almost same proof as that of Lemma C.5, we can deduce the following lemma.

Lemma C.6. *Fix scalars $\rho \in (0, 1)$ and $v \in (0, \infty)$. For brevity let $K := 2A^H$. If $\Delta \le 0.3$, for any measurable function $f : \mathcal{A}(\mathcal{X})^{TH} \times \{0, 1\}^{TH} \to \{0, 1\}$ and $k \in [K]$, it holds that*

$$\mathbb{P}_k \left(f(H_{T+1}) = 1 \right) > \exp\left(-s - v\right) \left(\mathbb{P}_0 \left(f(H_{T+1}) = 1 \right) - \rho - \frac{4\Delta^2}{v} n_k \right),$$

where $s := \dfrac{4\Delta}{3} \log \dfrac{1}{\rho} + \sqrt{2v \log \dfrac{1}{\rho}}$ *and* $n_k := \mathbb{E}_0 \left[\sum_{t=1}^{T} \mathbb{I}_{\left\{ x_H^t = x_k^* a_H^t = a_k^* \right\}} \right]$.

Regret Lower Bound. Now we are ready to prove the regret lower bound. For each $k \in [K]$, let E_k be an event $\sum_{t=1}^{T} \mu_{1:H}^t(x_k^*, a_k^*) \le 0.5T$. Note that while implicit, $\sum_{t=1}^{T} \mu_{1:H}^t(x_k^*, a_k^*)$ is a function of H_{T+1}, and thus, E_k can be rewritten in a form of $f(H_{T+1}) = 1$. As we verify later, our choice of Δ is smaller than 0.3. Therefore from the previous lemma,

$$\max_{k \in [K]} \mathbb{P}_k(E_k) \ge \frac{1}{K} \sum_{k \in [K]} \mathbb{P}_k(E_k) > \exp\left(-s - v\right) \left(\frac{1}{K} \sum_{k=1}^{K} \mathbb{P}_0 \left(E_k \right) - \rho - \frac{4T\Delta^2}{Kv} \right).$$

Note that $\sum_{k \in [K]} \mathbb{P}_0 \left(E_k \right) = \mathbb{E}_0 [\sum_{k \in [K]} \mathbb{I}_{\{E_k\}}] \ge K - A^{H-1}$. To see why it is the case, suppose that $\mathbb{I}_{\{E_k\}} = 0$ for some k, and let $(x_h^*, a_h^*)_{h=1}^{H}$ be information sets and actions from the root to (x_k^*, a_k^*). Then $\sum_{t=1}^{T} \mu_{1:H}^t(x_k^*, a_k^*) > 0.5T$ means that $\sum_{t=1}^{T} \mu_{1:h}^t(x_h^*, a_h^*) > 0.5T$ for all h since $\mu_{1:h}^t(x_h^*, a_h^*) = \mu_{1:h-1}^t(x_{h-1}^*, a_{h-1}^*) \mu_h^t(a_h^*|x_h^*)$. Furthermore it also holds that $\sum_{t=1}^{T} \mu_h^t(a_h^*|x_h^*) > 0.5T$ for any h. If there is no branch along the path to the leaf action node to the root node, every trajectory to leaf action nodes certainly contains x_1^*. Therefore $\sum_{t=1}^{T} \mu_1^t(a_1|x_1^*) \le 0.5T$ holds for any $a_1 \in \mathcal{A} \setminus \{a_1^*\}$, and $\mathbb{I}_{\{E_k\}} = 0$ for some k implies that $\mathbb{I}_{\{E_j\}} = 1$ for some $j \ne k$. If there is an additional branch, the same argument holds for all leaf action nodes except those in the branch. In the worst case, the number of leaf action nodes in a branch is A^{H-1}. Thus,

$$\max_{k \in [K]} \mathbb{P}_k(E_k) > \exp\left(-s - v\right) \left(1 - \frac{1}{A} - \rho - \frac{4T\Delta^2}{Kv} \right).$$

Setting $v = \dfrac{32T\Delta^2}{K}$ and $\rho = \dfrac{1}{8}$,

$$\max_{k \in [K]} \mathbb{P}_k(E_k) > \frac{1}{4} \exp\left(-4\Delta \log 2 - 8\Delta \sqrt{\frac{3T}{K} \log 2} - \frac{32T\Delta^2}{K} \right),$$

where we used an assumption that $A \ge 2$. Equating the right hand side to δ and solving for Δ, we choose

$$\Delta = \sqrt{\left(\frac{K}{16T} \log 2 + \frac{1}{8} \sqrt{\frac{3K}{T} \log 2} \right)^2 + \frac{K}{32T} \log \frac{1}{4\delta}} - \frac{K}{16T} \log 2 - \frac{1}{8} \sqrt{\frac{3K}{T} \log 2} < \sqrt{\frac{K}{32T} \log \frac{1}{4\delta}} < 0.3, \quad (3)$$

where the middle inequality follows since $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$ for $a, b \in (0, \infty)$. Then it holds that $\max_{k \in [K]} \mathbb{P}_k(E_k) > \delta$. Because the event E_k implies that the regret is greater than or equal to $0.25 T \Delta H$ in k-th game, we have a lower bound $\Omega(\sqrt{-TA^H \log(4\delta)}) = \Omega(\sqrt{-TXA \log(4\delta)})$, where $X \leq 2(1 + A + \cdots + A^{H-1}) = X_H(1 + 1/A + \cdots + 1/A^{H-1}) \leq 2X_H$ is used.

Sample Complexity Lower Bound. Now we are ready to prove the sample complexity lower bound. Let E_k be an event $\mu_1^{T+1}(x_k^*, a_k^*) \leq 0.5$ for each $k \in [K]$. By the same argument as the one in the regret lower bound proof, $\max_{k \in [K]} \mathbb{P}_k(E_k) > \delta$ for Δ in Equation 3. Since in k-th game, the event E_k implies that the simple regret of μ^{T+1} is greater than or equal to $0.25\Delta H$, we have a lower bound for the simple regret of $\Omega(\sqrt{-XA/T \log(4\delta)})$. Therefore unless $T \geq -XA/\varepsilon^2 \log(4\delta)$, the algorithm fails to output ε-optimal policy with probability greater than δ.

D. General tools

All proofs of the following sections will focus on the max player. We will denote by $\mathcal{A}(\mathcal{X})^t$ its history at episode t and use $(\mathcal{F}_t)_{t \in [\![0,T]\!]}$ the filtration such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_t = \sigma\left(\nu^1, r^1, \mathcal{A}(\mathcal{X})^1 \ldots \nu^t, r^t, \mathcal{A}(\mathcal{X})^t, \nu^{t+1}\right)$ and $\mathcal{F}_T = \sigma\left(\nu^1, r^1, \mathcal{A}(\mathcal{X})^1, \ldots, \nu^T, r^T \mathcal{A}(\mathcal{X})^T\right)$ is the filtration of σ-algebra generated by the relevant random variables for the max-player up to the round $t+1$.

We first provide a slight generalization of Lemma 1 by Neu (2015) for our settings, as γ^t is not fixed in advance in the adaptive case.

Lemma D.1. *Let $h \in [H]$, $(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)$, $\tau_h(x_h, a_h)$ a stopping time with respect to $(\mathcal{F}_t)_{t \in [\![0,T]\!]}$, and $\gamma'_h(x_h, a_h) > 0$ a fixed constant such that for all $t \leq \tau_h(x_h, a_h)$, $\gamma_h^t(x_h, a_h) \geq \gamma'_h(x_h, a_h)$. Then, with probability $1 - \delta'$,*

$$\widetilde{L}_h^{\tau_h(x_h,a_h)}(x_h, a_h) - L_h^{\tau_h(x_h,a_h)}(x_h, a_h) \leq \frac{\log(1/\delta')}{2\gamma'_h(x_h, a_h)} .$$

Proof. We first define the random process $(S^t)_{t \in [\![0,t]\!]}$ with respect to $(\mathcal{F}_t)_{t \in [\![0,T]\!]}$

$$S^t := \exp\left[2\gamma'_h(x_h, a_h) \sum_{k=1}^{t \wedge \tau_h(x_h,a_h)} \left(\widetilde{\ell}_h^k(x_h, a_h) - \ell_h^k(x_h, a_h) \right) \right] .$$

With the inequality $\frac{z}{1+z/2} \leq \log(1+z)$ for $z \geq 0$, we have for all $t \leq \tau_h(x_h, a_h)$, with $\overline{\mu}_{1:h}^t := \mu_{1:h}^t(x_h, a_h)$,

$$2\gamma'_h(x_h, a_h)\widetilde{\ell}_h^t(x_h, a_h) = 2\gamma'_h(x_h, a_h) \frac{1 - r_h^t}{\overline{\mu}_{1:h}^t + \gamma_h^t(x_h, a_h)} \mathbb{I}_{\{x_h = x_h^t, a_h = a_h^t\}}$$

$$\leq 2\gamma'_h(x_h, a_h) \frac{1 - r_h^t}{\overline{\mu}_{1:h}^t + \gamma'_h(x_h, a_h)(1 - r_h^t)} \mathbb{I}_{\{x_h = x_h^t, a_h = a_h^t\}}$$

$$= \frac{2\gamma'_h(x_h, a_h)(1 - r_h^t)/\overline{\mu}_{1:h}^t}{1 + \gamma'_h(x_h, a_h)(1 - r_h^t)/\overline{\mu}_{1:h}^t} \mathbb{I}_{\{x_h = x_h^t, a_h = a_h^t\}}$$

$$\leq \log(1 + 2\gamma'_h(x_h, a_h)\widehat{\ell}_h^t(x_h, a_h)) .$$

Which gives, using $1 + z \leq e^z$,

$$\mathbb{E}\left[\exp\left(2\gamma'_h(x_h, a_h)\widetilde{\ell}_h^t(x_h, a_h) \right) | \mathcal{F}_{t-1} \right] \leq \mathbb{E}\left[1 + 2\gamma'_h(x_h, a_h)\widehat{\ell}_h^t(x_h, a_h) | \mathcal{F}_{t-1} \right]$$

$$= 1 + 2\gamma'_h(x_h, a_h)\ell_h^t(x_h, a_h)$$

$$\leq \exp\left(2\gamma'_h(x_h, a_h)\ell_h^t(x_h, a_h) \right) .$$

We thus have that $(S^t)_{t \in [\![0,T]\!]}$ is a super-martingale, and using the Markov inequality we get

$$\mathbb{P}\left(\widetilde{L}_h^{\tau_h(x_h,a_h)}(x_h,a_h) - L_h^{\tau_h(x_h,a_h)}(x_h,a_h) \geq \log(1/\delta')/\left(2\gamma_h'(x_h,a_h)\right)\right) = \mathbb{P}\left(S^T \geq 1/\delta'\right)$$
$$\leq \mathbb{E}\left(S^T\right)\delta'$$
$$\leq \delta'.$$

$\square$

We also state a simple consequence of Azuma-Hoeffding inequality that we will use multiple times in the following sections.

Lemma D.2. *Let $(u_t)_{t\in[T]}$ be a random process adapted to $(\mathcal{F}_t)_{t\in[T]}$ such that $0 \leq u_t \leq H$ for all $t \in [T]$. Then, with a probability at least $1 - \delta'$,*

$$\sum_{t=1}^{T}\left[u_t - \mathbb{E}\left(u_t|\mathcal{F}_{t-1}\right)\right] \leq H\sqrt{2T\log(1/\delta')}.$$

The same property can also be shown for $\sum_{t=1}^{T}\left[\mathbb{E}\left(u_t|\mathcal{F}_{t-1}\right) - u_t\right]$.

Proof. We define the martingale $(M_t)_{t\in[[0,T]]}$ adapted to $(\mathcal{F}_t)_{t\in[[0,T]]}$ with

$$M_t := \sum_{k=1}^{t}\left[u_k - \mathbb{E}\left(u_k|\mathcal{F}_{k-1}\right)\right].$$

We have, thanks to the hypothesis, for all $1 \leq t \leq T$, that $-H \leq M_t - M_{t-1} \leq H$. Azuma-Hoeffding inequality then allows us to conclude

$$\mathbb{P}\left(M_T \geq H\sqrt{2T\log(1/\delta')}\right) \leq \exp\left(-2\frac{2H^2T\log(1/\delta')}{4H^2T}\right) = \delta'.$$

The proof for $\sum_{t=1}^{T}\left[\mathbb{E}\left(u_t|\mathcal{F}_{t-1}\right) - u_t\right]$ is exactly the same. $\square$

Finally, we state a lemma useful for the analysis of `Adaptive FTRL`.

Lemma D.3. *Let $(u_t)_{1\leq t\leq T}$ be a non-negative sequence that verifies for all $t \in [T]$, $u_t \leq 1 + U_{t-1}$ where $U_t = \sum_{k=1}^{t}u_k$. Then*

$$\sum_{t=1}^{T}\frac{u_t}{1+U_{t-1}} \leq \log_2(1+U_T).$$

Proof. Let $\lambda_t = \frac{u_t}{1+U_{t-1}}$. Using the concavity of $\log_2$, we get $\lambda_t \leq \log_2(1+\lambda_t)$ as $\lambda_t \in [0,1]$ by assumption. Summing over t then yields

$$\sum_{t=1}^{T}\lambda_t \leq \sum_{t=1}^{T}\log_2(1+\lambda_t) = \sum_{t=1}^{T}\log_2\left(\frac{1+U_t}{1+U_{t-1}}\right) = \log_2(1+U_T).$$

$\square$

E. Proof of `Balanced FTRL`

Space of realization plans Let $\Pi_{\max}^{1:} := \{\mu_{1:}, \mu \in \Pi_{\max}\}$ the set of max-player realization plans. We can notice that this space is a restriction of $\mathbb{R}_{\geq 0}^{A_\mathcal{X}}$ to an affine subspace of $\mathbb{R}^{A_\mathcal{X}}$, defined by the X constraints, for each $x_h \in \mathcal{X}$,

$$\sum_{a_h\in\mathcal{A}(x_h)}\mu_{1:h}(x_h,a_h) = \mu_{1:h-1}(x_{h-1},a_{h-1}),$$

23

with the convention $\mu_{1:0}(x_0, a_0) = 1$. We thus write $\Pi_{\max}^{1:} = (F + u) \cap \mathbb{R}_{\geq 0}^{A_{\mathcal{X}}}$, with F a linear subspace of $\mathbb{R}^{A_{\mathcal{X}}}$ and $u \in \Pi_{\max}^{1:}$.

Average profile We especially get that the average $\frac{1}{T} \sum_{t=1}^{T} \mu_{1:}^t$ is also in $\Pi_{\max}^{1:}$. We define an average realization plan $\overline{\mu}$ of the max-player as one of the $\overline{\mu} \in \Pi_{\max}$ that verifies

$$\overline{\mu}_{1:} = \frac{1}{T} \sum_{t=1}^{T} \mu_{1:}^t .$$

Doing the same step for $\overline{\nu}$, an average realization of the min-player, we get the existence of an average profile $(\overline{\mu}, \overline{\nu})$. Note that if the average profile is positive over all information states, then it is unique. It has the same properties as the one defined in Kozuno et al. (2021).

Global regularizer We will assume that the Ψ_h functions are supported on $\mathbb{R}_{\geq 0}^{A(\mathcal{X}_h)}$, and denote by $\Psi : \mathbb{R}_{\geq 0}^{A_{\mathcal{X}}} \to \mathbb{R}$ the function $\Psi(\mu_{1:}) = \sum_{h=1}^{H} \frac{1}{\eta_h} \Psi_h(\mu_{1:h} \cdot p_{1:h}^\star)$. This function is strictly convex and non-positive when using either Tsallis or Shannon entropy as Ψ_h function. Its definition with the coordinates of the realization plan will be important as we will consider its gradient. The associated Bregman divergence is

$$\mathcal{D}_\Psi(\mu_{1:}^1, \mu_{1:}^2) := \Psi(\mu_{1:}^1) - \Psi(\mu_{1:}^2) - \langle \nabla\Psi(\mu_{1:}^2), \mu_{1:}^1 - \mu_{1:}^2 \rangle .$$

We first give a lemma that justifies the use of the balanced transitions.

Lemma E.1. *For any $\nu \in \Pi_{\min}$, we have*

$$\sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} A_{x_h} \frac{p_{1:h}^\nu(x_h)}{p_{1:h}^\star(x_h)} = A_{\mathcal{X}} .$$

Furthermore, when the size of the action set is fixed to A, we also have for any $h \in [H]$

$$\sum_{x_h \in \mathcal{X}_h} A_{x_h} \frac{p_{1:h}^\nu(x_h)}{p_{1:h}^\star(x_h)} \leq A^{h-H} A_{\mathcal{X}} .$$

Proof. We first show by induction on h that that for all $x_h \in \mathcal{X}_h$,

$$\sum_{h'=1}^{h-1} \sum_{x_{h'} \in \mathcal{X}_{h'}} A_{x_{h'}} \frac{p_{1:h'}^\nu(x_{h'})}{p_{1:h'}^\star(x_{h'})} + \sum_{x_h \in \mathcal{X}_h} A_{x_h}^\tau \frac{p_{1:h}^\nu(x_h)}{p_{1:h}^\star(x_h)} = A_{\mathcal{X}} . \tag{4}$$

For $h = 1$, we have

$$\sum_{x_1 \in \mathcal{X}_1} A_{x_1}^\tau \frac{p_{1:1}^\nu(x_1)}{p_{1:1}^\star(x_1)} = A_{\mathcal{X}} \sum_{x_1 \in \mathcal{X}_1} p_0(x_1) = A_{\mathcal{X}} .$$

Assuming the property holds at step h, to obtain the property at step $h+1$ we use

$$\sum_{x_{h+1} \in \mathcal{X}_{h+1}} A_{x_{h+1}}^\tau \frac{p_{1:h+1}^\nu(x_{h+1})}{p_{1:h+1}^\star(x_{h+1})} = \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \sum_{(\ldots, x_h, a_h, x_{h+1})} A_{x_{h+1}}^\tau \frac{p_{1:h+1}^\nu(x_{h+1})}{p_{1:h}^\star(x_h) p_h^\star(x_{h+1}|x_h, a_h)}$$

$$= \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \sum_{(\ldots, x_h, a_h, x_{h+1})} \frac{p_{1:h+1}^\nu(x_{h+1})}{p_{1:h}^\star(x_h)} \sum_{(\ldots, x_h, a_h, x_{h+1}')} A_{x_{h+1}'}^\tau$$

$$= \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{p_{1:h}^\nu(x_h)}{p_{1:h}^\star(x_h)} \sum_{(\ldots, x_h, a_h, x_{h+1}')} A_{x_{h+1}'}^\tau$$

$$= \sum_{x_h \in \mathcal{X}_h} \frac{p_{1:h}^\nu(x_h)}{p_{1:h}^\star(x_h)} (A_{x_h}^\tau - A_{x_h}) .$$

24

As for any $x_H \in \mathcal{X}_H$, $A^\tau_{x_H} = A_{x_H}$ (as no state follows a final state), we obtain the first equality using (4) with $h = H$.

For the inequality, we can notice that, if the size of the action set is fixed equal to A, we have for any $h \in [H-1]$:

$$A^\tau_{x_h} = \sum_{x' \geq x_h} A(x') \geq \sum_{a_h \in \mathcal{A}(x_h)} \sum_{(\dots, x_h, a_h, \dots x_H)} A_{x_H}$$

$$= A \sum_{a_h \in \mathcal{A}(x_h)} \text{Card}\{x_H \in \mathcal{X}_H | (\dots, x_h, a_h, \dots x_H)\} \geq A \sum_{a_h \in \mathcal{A}(x_h)} A^{H-h-1} = A_{x_h} A^{H-h}.$$

The inequality $A^\tau_{x_h} \geq A_{x_h} A^{H-h}$ also trivially holds for $h = H$ as explained before. Using again (4), this time with any $h \in [H]$, we obtain

$$A_\mathcal{X} = \sum_{h'=1}^{h-1} \sum_{x_{h'} \in \mathcal{X}_{h'}} A_{x_{h'}} \frac{p^\nu_{1:h'}(x_{h'})}{p^\star_{1:h'}(x_{h'})} + \sum_{x_h \in \mathcal{X}_h} A^\tau_{x_h} \frac{p^\nu_{1:h}(x_h)}{p^\star_{1:h}(x_h)}$$

$$\geq \sum_{x_h \in \mathcal{X}_h} A^\tau_{x_h} \frac{p^\nu_{1:h}(x_h)}{p^\star_{1:h}(x_h)}$$

$$\geq A^{H-h} \sum_{x_h \in \mathcal{X}_h} A_{x_h} \frac{p^\nu_{1:h}(x_h)}{p^\star_{1:h}(x_h)}.$$

Which concludes the proof. $\qquad\square$

Lemma E.2. *Assume that the size of the action set is fixed equal to $A \geq 2$, then $H \leq \sqrt{A_\mathcal{X}}$ and we have the inequality:*

$$\sum_{h=1}^{H} A^{(h-H)/2} \leq 2 + \sqrt{2}.$$

Proof. The first inequality is simply obtained with

$$A_\mathcal{X} \geq \sum_{h=1}^{H} A^h \geq \sum_{h=1}^{H} 2^h \geq (2^{H+1} - 2) \geq H^2$$

where the first inequality comes from the fact that, at each time step h, the player remembers its $h-1$ past actions, and the last comes from the fact that $(2^{H+1} - 2)/H^2$ is increasing for $H \geq 3$ and is more than 1 for $H = 1, 2, 3$.

For the second inequality we use

$$\sum_{h=1}^{H} A^{(h-H)/2} \leq \sum_{h=1}^{H} 2^{(h-H)/2} \leq \sum_{h=0}^{+\infty} A^{h/2} \leq 1/(1 - 1/\sqrt{2}) = 2 + 2\sqrt{2}.$$

$\qquad\square$

We now try to bound the regret. We first decompose it, using the same decomposition as in Zimmert & Seldin (2019) for the expected part.

Lemma E.3. *The regret of* `Balanced FTRL` *can be decomposed into*

$$\mathfrak{R}_T \leq \underbrace{\sum_{t=1}^{T} \left\langle \mu^t_{1:}, \ell^t - \widetilde{\ell}^t \right\rangle}_{BIAS\ I} + \underbrace{\max_{\mu^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \left\langle \mu^\dagger_{1:}, \widetilde{\ell}^t - \ell^t \right\rangle}_{BIAS\ II}$$

$$+ \underbrace{\max_{\mu \in \Pi_{\max}} [-\Psi(\mu_{1:})]}_{REG} + \underbrace{\sum_{t=1}^{T} \mathcal{D}_{\Psi^\star} \left(\nabla\Psi(\mu^t_{1:}) - \widetilde{\ell}^t, \nabla\Psi(\mu^t_{1:}) \right)}_{VAR}.$$

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

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

where $\Phi(y) = \sup_{\mu \in \Pi_{\max}} \langle \mu_{1:}, y \rangle - \Psi(\mu_{1:})$.

Summing the first two terms over t gives the two BIAS terms. For the REG term, summing over t yields, by telescoping,

$$\sum_{t=1}^T \left[\Phi\left(-\widetilde{L}^{t-1}\right) - \Phi\left(-\widetilde{L}^t\right) - \left\langle \mu_{1:}^\dagger, \widetilde{\ell}^t \right\rangle \right] = \Phi(0) - \Phi(-\widetilde{L}^t) - \left\langle \mu_{1:}^\dagger, \widetilde{L}^t \right\rangle$$
$$\leq \max_{\mu \in \Pi_{\max}} \left[-\Psi(\mu_{1:}) \right] + \Psi(\mu_{1:}^\dagger)$$
$$\leq \max_{\mu \in \Pi_{\max}} \left[-\Psi(\mu_{1:}) \right]$$

as $\mu_{1:}^\dagger \in \Pi_{\max}^{1:}$ for the first inequality, and Ψ is a non-positive function for the second inequality.

We then define $\Psi^\star$ the convex conjugate of Ψ, with $\Psi^\star(y) = \sup_{x \in \mathbb{R}_{\geq 0}^{A\mathcal{X}}} \langle x, y \rangle - \Psi(x)$, and, thanks to the decomposition $\Pi_{\max}^{1:} = (F + u) \cap \mathbb{R}_{\geq 0}^{A\mathcal{X}}$, get

$$\left\langle \mu_{1:}^t, \widetilde{\ell}^t \right\rangle + \Phi\left(-\widetilde{L}^t\right) - \Phi\left(-\widetilde{L}^{t-1}\right)$$
$$= \left\langle \mu_{1:}^t, \widetilde{\ell}^t \right\rangle + \Phi\left(\nabla\Psi(\mu_{1:}^t) + g_t - \widetilde{\ell}^t\right) - \Phi\left(\nabla\Psi(\mu_{1:}^t) + g_t\right) \tag{1}$$
$$= \left\langle \mu_{1:}^t, \widetilde{\ell}_t \right\rangle + \Phi\left(\nabla\Psi(\mu_{1:}^t) - \widetilde{\ell}^t\right) - \Phi\left(\nabla\Psi(\mu_{1:}^t)\right) \tag{2}$$
$$\leq \left\langle \mu_{1:}^t, \widetilde{\ell}_t \right\rangle + \Psi^*\left(\nabla\Psi(\mu_{1:}^t) - \widetilde{\ell}^t\right) - \Psi^*\left(\nabla\Psi(\mu_{1:}^t)\right) \tag{3}$$
$$= \mathcal{D}_{\Psi^\star}\left(\nabla\Psi(\mu_{1:}^t) - \widetilde{\ell}^t, \nabla\Psi(\mu_{1:}^t)\right) .$$

Where we used successively:

(1) As $\mu^t = \arg\min_{\mu \in \Pi_{\max}} \left\langle \mu_{1:}, \widetilde{L}^{t-1} \right\rangle + \Psi(\mu_{1:})$, we have

$$\widetilde{L}^{t-1} + \nabla\Psi\left(\mu_{1:}^t\right) + g^t = 0$$

where $g^t \in F^\perp$.

(2) For all $y \in \mathbb{R}^{A\mathcal{X}}$, $\Phi\left(y + g^t\right) = \Phi\left(y\right) + \langle u, g^t \rangle$.

(3) Because Φ is a constrained version over $\Pi_{\max}^{1:}$ of Ψ^*, $\Phi \leq \Psi^*$. And as Ψ is convex, $\mu_{1:}^t = \arg\max_{x \in \mathbb{R}_{\geq 0}^{A\mathcal{X}}} \langle x, \nabla\Psi(\mu_{1:}^t) \rangle - \Psi(x)$, which implies that the maxima associated to $\Psi^*(\nabla\Psi(\mu_{1:}^t))$ is reached on $\Pi_{\max}^{1:}$. $\qquad\square$

We first give upper bounds on the BIAS terms when using the IX estimations.

Lemma E.4. *For any sequence $\gamma_h = \gamma > 0$, with probability at least $1 - 2\delta/3$, we have*

$$BIAS\ I \leq H\sqrt{2T\iota} + \gamma A_{\mathcal{X}}T \qquad and \qquad BIAS\ II \leq H\frac{\iota}{2\gamma}\,.$$

If the size of the action set is fixed and $\gamma_h = A^{(H-h)/2}\gamma$ instead, with probability at least $1 - 2\delta/3$, we have

$$BIAS\ I \leq H\sqrt{2T\iota} + (2+\sqrt{2})\gamma A_{\mathcal{X}}T \qquad and \qquad BIAS\ II \leq (1+1/\sqrt{2})\frac{\iota}{\gamma}\,.$$

Proof. We can first decompose BIAS I into

$$\text{BIAS I} = \sum_{t=1}^{T}\left\langle \mu_{1:}^t, \mathbb{E}[\widetilde{\ell}^t|\mathcal{F}_{t-1}] - \widetilde{\ell}^t\right\rangle + \sum_{t=1}^{T}\left\langle \mu_{1:}^t, \ell^t - \mathbb{E}[\widetilde{\ell}^t|\mathcal{F}_{t-1}]\right\rangle\,.$$

Using Lemma D.2, as $\left\langle \mu_{1:}^t, \widetilde{\ell}^t\right\rangle \leq H$, we can bound the first term with probability $1 - \delta/3$ by

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

The second term can be upper-bounded by using the Lemma E.1 and $\gamma_h = \gamma$ for the last equality,

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle \mu_{1:}^t, l^t - \mathbb{E}[\widetilde{\ell}^t|\mathcal{F}_{t-1}]\right\rangle &\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} p_{1:h}^{\nu^t}(x_h)\mu_{1:h}^t(x_h,a_h)\left(1 - \frac{\mu_{1:h}^t(x_h,a_h)}{\mu_{1:h}^t(x_h,a_h) + \gamma_h^\star(x_h,a_h)}\right) \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} p_{1:h}^{\nu^t}(x_h)\frac{\mu_{1:h}^t(x_h,a_h)\gamma_h(x_h,a_h)}{\mu_{1:h}^t(x_h,a_h) + \gamma_h^\star(x_h,a_h)} \\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} p_{1:h}^{\nu^t}(x_h)\gamma_h^\star(x_h,a_h) \\
&= \sum_{t=1}^{T}\sum_{h=1}^{H}\gamma_h\sum_{x_h\in\mathcal{X}_h} A_{x_h}\frac{p_{1:h}^{\nu^t}(x_h)}{p_{1:h}^\star(x_h)} \\
&= \gamma T A_{\mathcal{X}}\,.
\end{aligned}$$

In the setting with a fixed size of the action set, using the depth-wise inequality of Lemma E.1, and Lemma E.2 for the last inequality, we can get instead

$$\begin{aligned}
\sum_{t=1}^{T}\left\langle \mu_{1:}^t, l^t - \mathbb{E}[\widetilde{\ell}^t|\mathcal{F}_{t-1}]\right\rangle &\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\gamma_h\sum_{x_h\in\mathcal{X}_h} A\frac{p_{1:h}^{\nu^t}(x_h)}{p_{1:h}^\star(x_h)} \\
&\leq \gamma A_{\mathcal{X}}\sum_{t=1}^{T}\sum_{h=1}^{H} A^{(H-h)/2}A^{h-H} \\
&= \gamma A_{\mathcal{X}}T\sum_{h=1}^{H} A^{(h-H)/2} \\
&\leq (2+\sqrt{2})\gamma A_{\mathcal{X}}T\,.
\end{aligned}$$

In order to bound BIAS II, we can use Lemma D.1. Indeed for all $(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)$, using $\delta' = \delta/(3A_{\mathcal{X}})$, the constant stopping time $\tau_h(x_h,a_h) = T$ and $\gamma_h'(x_h,a_h) = \gamma_h^\star(x_h,a_h)$, we have, as $\iota = \log(3A_{\mathcal{X}}/\delta)$, with probability at least $1 - \delta/(3A_{\mathcal{X}})$,

$$\widetilde{L}_h^T(x_h, a_h) - L_h^T(x_h, a_h) \le \frac{\iota}{2\gamma_h^\star(x_h, a_h)} \, .$$

By a union bound the inequality holds for all $(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)$ with probability at least $1 - \delta/3$. In this case, we get the bound, for any $\mu^\dagger \in \Pi_{\max}$, whith $\gamma_h = \gamma$

$$
\begin{aligned}
\sum_{t=1}^T \left\langle \mu_{1:}^\dagger, \widetilde{\ell}^t - \ell^t \right\rangle &= \sum_{h=1}^H \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \mu_h^\dagger(x_h, a_h) \left(\widetilde{L}_h^T(x_h, a_h) - L_h^T(x_h, a_h) \right) \\
&\le \iota \sum_{h=1}^H \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{\mu_h^\dagger(x_h, a_h)}{2\gamma_h^\star(x_h, a_h)} \\
&= \frac{\iota}{2} \sum_{h=1}^H \frac{1}{\gamma_h} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \mu_h^\dagger(x_h, a_h) p_{1:h}^\star(x_h) \\
&= \frac{\iota}{2} \sum_{h=1}^H \frac{1}{\gamma_h} \\
&= \frac{\iota}{2\gamma} H
\end{aligned}
$$

which gives the bound on BIAS II by taking the maximum over $\mu^\dagger \in \Pi_{\max}$. In the second setting we have instead, again with Lemma E.2

$$\sum_{t=1}^T \left\langle \mu_{1:}^\dagger, \widetilde{\ell}^t - \ell^t \right\rangle \le \frac{\iota}{2} \sum_{h=1}^H \frac{1}{\gamma_h} = \frac{\iota}{2\gamma} \sum_{h=1}^H A^{(h-H)/2} \le (1 + 1/\sqrt{2}) \frac{\iota}{\gamma} \, .$$

$\square$

Depending on the regularizer used in `Balanced FTRL`, we then state some upper bounds of the REG term.

Lemma E.5. *Assume that $\eta_h = \eta$, then with Tsallis entropy*

$$REG \le \frac{H^q}{\eta} A_{\mathcal{X}}^{1-q} \, ,$$

and with Shannon entropy

$$REG \le \frac{H}{\eta} \log(A_{\mathcal{X}}) \, .$$

Furthermore, if the size of the action set is fixed and $\eta_h = A^{(H-h)/2}\eta$ instead, we have with Shannon entropy

$$REG \le \frac{2 + \sqrt{2}}{\eta} \log(A_{\mathcal{X}}) \, .$$

Proof. The upper-bound with Tsallis entropy is a consequence of Hölder inequality, as for all $\mu_{1:} \in \Pi_{\max}^{1:}$,

$$
\begin{aligned}
-\Psi(\mu_{1:}) &= \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{1}{\eta_h} \left(p_{1:h}^{\star}(x_h) \mu_{1:h}(x_h, a_h) \right)^q \\
&= \frac{1}{\eta} \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \left(p_{1:h}^{\star}(x_h) \mu_{1:h}(x_h, a_h) \right)^q \\
&\leq \frac{1}{\eta} \left(\sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} p_{1:h}^{\star}(x_h) \mu_{1:h}(x_h, a_h) \right)^q \left(\sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} 1 \right)^{1-q} \\
&= \frac{H^q}{\eta} {A_{\mathcal{X}}}^{1-q} \,.
\end{aligned}
$$

When using Shannon entropy, we use that $p_{1:h}^{\star} \cdot \mu_{1:h}$ is a probability distribution on $\mathcal{A}(\mathcal{X}_h)$ for all $h \in [H]$, and get with Jensen inequality

$$
\begin{aligned}
-\Psi(\mu) &\leq \sum_{h=1}^{H} \frac{-1}{\eta_h} \Psi_h(p_{1:h}^{\star} \cdot \mu_{1:h}) \\
&\leq \sum_{h=1}^{H} \frac{1}{\eta_h} \log(A_{\mathcal{X}_h}) \\
&\leq \sum_{h=1}^{H} \frac{1}{\eta_h} \log(A_{\mathcal{X}}) \\
&= \frac{H}{\eta} \log(A_{\mathcal{X}}) \,.
\end{aligned}
$$

If the size of the action set is fixed and $\eta_h = A^{(H-h)/2}\eta$, we instead get with Lemma E.2

$$
-\Psi(\mu) \leq \sum_{h=1}^{H} \frac{1}{\eta_h} \log(A_{\mathcal{X}}) = \sum_{h=1}^{H} \frac{A^{(h-H)/2}}{\eta} \log(A_{\mathcal{X}}) \leq \frac{2+\sqrt{2}}{\eta} \log(A_{\mathcal{X}}) \,.
$$

$\square$

Upper bounding the VAR term is a little harder, but the idea is the same between the two regularizers, with each term of the sum over time being bounded separately in expectation.

Lemma E.6. *Let* $v_t = \mathcal{D}_{\Psi^\star}\left(\nabla\Psi(\mu_{1:}^t) - \widetilde{\ell}^t, \nabla\Psi(\mu_{1:}^t) \right)$ *for all* $t \in [T]$. *Then with a probability at least* $1 - \delta/3$,

$$
VAR \leq \sum_{t=1}^{T} \mathbb{E}[v_t | F_{t-1}] + H\sqrt{2T\iota} \,.
$$

Furthermore, if $\eta_h = \eta$, *we have with Tsallis entropy*

$$
\sum_{t=1}^{T} \mathbb{E}[v_t | F_{t-1}] \leq \eta \frac{T A_{\mathcal{X}}^q}{2q(1-q)} H^{1-q}
$$

and with Shannon entropy

$$
\sum_{t=1}^{T} \mathbb{E}[v_t | F_{t-1}] \leq \frac{\eta}{2} T A_{\mathcal{X}} \,.
$$

If $\eta_h = A^{(H-h)/2}\eta$ and the size of the action set is fixed to A instead, we then have with Shannon entropy

$$\sum_{t=1}^{T} \mathbb{E}[v_t|F_{t-1}] \leq (1 + 1/\sqrt{2})\eta T A_{\mathcal{X}}.$$

Proof. We can first notice, as Ψ is supported on $\mathbb{R}_{\geq 0}^{A_{\mathcal{X}}}$ and the estimated losses $\widetilde{\ell}^t$ are non-negative, that

$$v_t = \left\langle \mu_{1:}^t, \widetilde{\ell}^t \right\rangle + \Psi^\star\left(\nabla\Psi(\mu_{1:}^t) - \widetilde{\ell}^t\right) - \Psi^\star\left(\nabla\Psi(\mu_{1:}^t)\right) \leq \left\langle \mu_{1:}^t, \widetilde{\ell}^t \right\rangle \leq H.$$

Which lets us use Lemma D.2 to get, with a probability at least $1 - \delta/3$,

$$\sum_{t=1}^{T}\left(v_t - \mathbb{E}[v_t|\mathcal{F}_{t-1}]\right) \leq \sqrt{2T\log(3/\delta)} \leq \sqrt{2T\iota}.$$

And obtain the first inequality with the Doob decomposition of the random process $(\sum_{k=1}^{t} v_k)_{t\in[T]}$, as VAR $= \sum_{k=1}^{T} v_k$.

For the upper-bounds of $\sum_{t=1}^{T}\mathbb{E}[v_t|F_{t-1}]$, we define for all $t \in [T]$, $f_t(u) = \mathcal{D}_{\Psi^\star}\left(\nabla\Psi(\mu_{1:}^t) - u\widetilde{\ell}^t, \nabla\Psi(\mu_{1:}^t)\right)$ for $u \in [0,1]$, such that $f_t(0) = 0$ and $f_t(1) = v_t$. As for both entropy, $\Psi(\mu_{1:})$ (respectively $\Psi^\star(y)$) can be decomposed into $\Psi(\mu_{1:}) = \sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \Psi_{x_h,a_h}(\mu_{1:h}(x_h,a_h))$ (respectively $\Psi^\star(y) = \sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \Psi_{x_h,a_h}^\star(y(x_h,a_h)))$, the derivative of f_t can be expressed with

$$f_t'(u) = \sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \widetilde{\ell}_h^t(x_h,a_h)\left[\mu_{1:h}^t(x_h,a_h) - \nabla\Psi_{x_h,a_h}^\star\left(\nabla\Psi_{x_h,a_h}(\mu_{1:h}^t(x_h,a_h)) - u\widetilde{\ell}_h^t(x_h,a_h)\right)\right]. \quad (5)$$

When using Tsallis entropy, we have

$$\nabla\Psi_{x_h,a_h}(\mu_{1:h}(x_h,a_h)) = -\frac{q}{\eta_h}p_{1:h}^\star(x_h)^q\mu_{1:h}(x_h,a_h)^{q-1}$$

$$\nabla\Psi_{x_h,a_h}^\star(y(x_h,a_h)) = \left(-\frac{\eta_h y(x_h,a_h)}{qp_{1:h}^\star(x_h)^q}\right)^{\frac{-1}{1-q}}$$

which yields

$$\nabla\Psi_{x_h,a_h}^\star\left(\nabla\Psi_{x_h,a_h}(\mu_{1:h}^t(x_h,a_h)) - u\widetilde{\ell}_h^t(x_h,a_h)\right)$$

$$= \left[\frac{-\eta_h}{qp_{1:h}^\star(x_h)^q}\left(\frac{-q}{\eta_h}p_{1:h}^\star(x_h)\mu_{1:h}^t(x_h,a_h)^{q-1} - u\widetilde{\ell}_h^t(x_h,a_h)\right)\right]^{\frac{-1}{1-q}}$$

$$= \mu_{1:h}^t(x_h,a_h)\left[1 - u\frac{\eta_h\widetilde{\ell}_h^t(x_h,a_h)}{qp_{1:h}^\star(x_h)^q}\mu_{1:h}^t(x_h,a_h)^{1-q}\right]^{\frac{-1}{1-q}}$$

$$\geq \mu_{1:h}^t(x_h,a_h)\left[1 - u\frac{\eta_h\widetilde{\ell}_h^t(x_h,a_h)}{q(1-q)p_{1:h}^\star(x_h)^q}\mu_{1:h}^t(x_h,a_h)^{1-q}\right]$$

where we used $(1+y)^{\frac{-1}{1-q}} \geq 1 - \frac{y}{1-q}$ for $y \geq 0$. Plugging this in (5), along with $\widetilde{\ell}_h^t(x_h,a_h)\mu_{1:h}^t(x_h,a_h) \leq \mathbb{I}_{\{x_h^t=x_h,a_h^t=a_h\}}$, gives

30

$$f'_t(u) \le u \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \widetilde{\ell}^t_h(x_h,a_h)\mu^t_{1:h}(x_h,a_h)\frac{\eta_h \widetilde{\ell}^t_h(x_h,a_h)}{q(1-q)p^\star_{1:h}(x_h)^q}\mu^t_{1:h}(x_h,a_h)^{1-q}$$

$$\le u \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \mathbb{I}_{\{x^t_h=x_h,a^t_h=a_h\}}\frac{\eta_h}{q(1-q)p^\star_{1:h}(x_h)^q}\mu^t_{1:h}(x_h,a_h)^{-q}\,.$$

Summing as u goes from 0 to 1 and conditioning yields

$$\mathbb{E}[v_t|\mathcal{F}_{t-1}] \le \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \mathbb{E}\left[\mathbb{I}_{\{x^t_h=x_h,a^t_h=a_h\}}\frac{\eta_h}{2q(1-q)p^\star_{1:h}(x_h)^q}\mu^t_{1:h}(x_h,a_h)^{-q}\Big|\mathcal{F}_{t-1}\right]$$

$$= \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \frac{\eta_h p^{\nu^t}_{1:h}(x_h)}{2q(1-q)p^\star_{1:h}(x_h)^q}\mu^t_{1:h}(x_h,a_h)^{1-q}$$

$$\le \frac{\eta}{2q(1-q)}\left(\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\frac{p^{\nu^t}_{1:h}(x_h)}{p^\star_{1:h}(x_h)}\right)^q \left(\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}p^{\nu^t}_{1:h}(x_h)\mu^t_{1:h}(x_h,a_h)\right)^{1-q}$$

$$= \frac{\eta H^{1-q}}{2q(1-q)}A_\mathcal{X}{}^q$$

where we used $\eta_h = \eta$, the Hölder inequality, and Lemma E.1 for the last equality.

When using Shannon entropy, we have

$$\nabla\Psi_{x_h,a_h}(\mu_{1:h}(x_h,a_h)) = \frac{p^\star_{1:h}(x_h)}{\eta_h}\left[\log(p^\star_{1:h}(x_h)\mu_{1:h}(x_h,a_h))-1\right]$$

$$\nabla\Psi^\star_{x_h,a_h}(y(x_h,a_h)) = \exp\left[\frac{\eta_h}{p^\star_{1:h}(x_h)}\left(y(x_h,a_h)\right)+1-\log(p^\star_{1:h}(x_h))\right]$$

and

$$\nabla\Psi^\star_{x_h,a_h}\left(\nabla\Psi_{x_h,a_h}(\mu^t_{1:h}(x_h,a_h))-u\widetilde{\ell}^t_h(x_h,a_h)\right)$$

$$= \exp\left[\frac{\eta_h}{p^\star_{1:h}(x_h)}\left(\frac{p^\star_{1:h}(x_h)}{\eta_h}\log(p^\star_{1:h}(x_h)\mu^t_{1:h}(x_h,a_h))-u\widetilde{\ell}^t_h(x_h,a_h)\right)-\log(p^\star_{1:h}(x_h))\right]$$

$$= \mu^t_{1:h}(x_h,a_h)\exp\left[-u\frac{\eta_h\widetilde{\ell}^t_h(x_h,a_h)}{p^\star_{1:h}(x_h)}\right]$$

$$\ge \mu^t_{1:h}(x_h,a_h)\left[1-u\frac{\eta_h\widetilde{\ell}^t_h(x_h,a_h)}{p^\star_{1:h}(x_h)}\right]\,.$$

Using this time $1-y \ge \exp(-y)$ for all $y \ge 0$. With (5), this gives

$$h'_t(u) \le u \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \widetilde{\ell}^t_h(x_h,a_h)\mu^t_{1:h}(x_h,a_h)\frac{\eta_h\widetilde{\ell}^t_h(x_h,a_h)}{p^\star_{1:h}(x_h)}$$

$$\le u \sum_{h=1}^{H} \sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)} \mathbb{I}_{\{x^t_h=x_h,a^t_h=a_h\}}\frac{\eta_h}{p^\star_{1:h}(x_h)\mu^t_{1:h}(x_h,a_h)}\,.$$

31

Again, summing from 0 to 1, using $\eta_h = \eta$ and Lemma E.1, leads to

$$\mathbb{E}[v_t|\mathcal{F}_{t-1}] \leq \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{\eta_h p_{1:h}^{\nu^t}(x_h, a_h)}{2p_{1:h}^{\star}(x_h)}$$

$$= \frac{\eta}{2} A_{\mathcal{X}}.$$

If the size of the action set is fixed and $\eta_h = A^{(H-h)/2}\eta$, we have instead, still with Shannon entropy,

$$\mathbb{E}[v_t|\mathcal{F}_{t-1}] \leq \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{\eta_h p_{1:h}^{\nu^t}(x_h, a_h)}{2p_{1:h}^{\star}(x_h)}$$

$$\leq \frac{\eta}{2} A_{\mathcal{X}} \sum_{h=1}^{H} A^{(h-H)/2}$$

$$\leq (1 + 1/\sqrt{2})\eta A_{\mathcal{X}}$$

where we used the depth-wise inequality of Lemma E.1 and Lemma E.2.

$\square$

We can now prove the main theorem of Section 4 using the previous lemmas.

Theorem 4.1. *Let $\delta \in (0, 1)$ and $\iota = \log(3A_{\mathcal{X}}/\delta)$. Fix the IX parameters to $\gamma_h = \sqrt{H\iota/(2A_{\mathcal{X}}T)}$ for all $h \in [H]$.*

Using `Balanced FTRL-Tsallis`*, $q \in (0, 1)$ and constant learning rates $\eta_h = \sqrt{2q(1-q)/T}\,(H/A_{\mathcal{X}})^{q-1/2}$, with probability at least $1 - \delta$,*

$$\mathfrak{R}_{\max}^T \leq \left(\sqrt{2/(q(1-q))} + 3\sqrt{2\iota}\right)\sqrt{HA_{\mathcal{X}}T}.$$

Using `Balanced FTRL-Shannon` *and constant learning rates $\eta_h = \sqrt{2H\log(A_{\mathcal{X}})/(A_{\mathcal{X}}T)}$, with probability at least $1 - \delta$,*

$$\mathfrak{R}_{\max}^T \leq \left(\sqrt{2\log(A_{\mathcal{X}})} + 3\sqrt{2\iota}\right)\sqrt{HA_{\mathcal{X}}T}.$$

Proof. From the decomposition of Lemma E.3, we have

$$\mathfrak{R}_{\max}^T \leq \text{BIAS I} + \text{BIAS II} + \text{REG} + \text{VAR}.$$

Lemma E.4 directly yields with probability at least $1 - 2\delta/3$

$$\text{BIAS I} + \text{BIAS II} \leq H\sqrt{2T\iota} + \gamma TA_{\mathcal{X}} + H\frac{\iota}{2\gamma}$$

that is minimized with $\gamma = \sqrt{H\iota/(2A_{\mathcal{X}}T)}$.

Then, using Tsallis entropy we have for the two other terms, from Lemma E.4 and Lemma E.6, with probability at least $1 - \delta/3$

$$\text{REG} + \text{VAR} \leq H\sqrt{2\iota T} + \frac{H^q}{\eta}A_{\mathcal{X}}^{1-q} + \eta\frac{TA_{\mathcal{X}}^q}{2q(1-q)}H^{1-q}$$

minimized with $\eta = \sqrt{2q(1-q)/T}\,H^{q-1/2}A_{\mathcal{X}}^{1/2-q}$. With the trivial upperbound $H \leq A_{\mathcal{X}}$, we then get by summing the two previous inequalities, with a probability at least $1 - \delta$

$$\mathfrak{R}_{\max}^T \leq 3\sqrt{2H\iota A_{\mathcal{X}}T} + \sqrt{\frac{2}{q(1-q)}HA_{\mathcal{X}}T}.$$

When using Shannon entropy, we instead have (with a probability at least $1 - \delta/3$) from the same lemmas:

$$\text{REG} + \text{VAR} \leq H\sqrt{2\iota T} + \frac{H}{\eta} \log(A_{\mathcal{X}}) + \frac{\eta}{2} A_{\mathcal{X}} T$$

minimized with $\eta = \sqrt{2H \log(A_{\mathcal{X}})(TA_{\mathcal{X}})}$, which in this case yields by summing, again with probability at least $1 - \delta$,

$$\mathfrak{R}_{\max}^T \leq 3\sqrt{2HTA_{\mathcal{X}}\iota} + \sqrt{2HTA_{\mathcal{X}}\log(A_{\mathcal{X}})}.$$

$\square$

In the special case of a fixed action set considered in the previous lemma, we can get a better rate for the regret of the max-player, here shown for Shannon entropy.

Theorem 4.4. *Assume that the size of the action set is constant equal to A and let $\delta \in (0, 1)$, $\iota = \log(3A_{\mathcal{X}}/\delta)$. Then,* `Balanced FTRL-Shannon` *with $\gamma_h = \sqrt{A^{H-h}\iota/(2A_{\mathcal{X}}T)}$ and $\eta_h = \sqrt{2A^{H-h}\log(A_{\mathcal{X}})/(A_{\mathcal{X}}T)}$ yields*

$$\mathfrak{R}_{\max}^T \leq \left(5\sqrt{\log(A_{\mathcal{X}})} + 8\iota\right)\sqrt{A_{\mathcal{X}}T}.$$

As in Corollary 4.3, this allows to compute an ε-optimal profile as long as long as

$$T \geq \mathcal{O}\left((A_{\mathcal{X}} + B_{\mathcal{Y}})\log(1/\delta + A_{\mathcal{X}} + B_{\mathcal{Y}})/\varepsilon^2\right).$$

Proof. The idea is the same as in the previous theorem, but using instead the upper bounds obtained with a fixed size of the action set that all holds with probability at least $1 - \delta$, as both η_h and γ_h parameters are in the form $\eta_h = A^{(H-h)/2}\eta$ and $\gamma_h = A^{(H-h)/2}\gamma$. The case $A = 1$ is ignored, as the regret is then trivially 0. With Lemma E.4, we now have:

$$\text{BIAS I} + \text{BIAS II} \leq H\sqrt{2T\iota} + (2 + \sqrt{2})\gamma TA_{\mathcal{X}} + (1 + 1/\sqrt{2})\frac{\iota}{\gamma}$$

minimized with $\gamma = \sqrt{\iota/(2TA_{\mathcal{X}})}$. The REG and VAR terms are in this case upper bounded, using Lemma E.5 and Lemma E.6, by

$$\text{REG} + \text{VAR} \leq H\sqrt{2T\iota} + \frac{2 + \sqrt{2}}{\eta}\log(A_{\mathcal{X}}) + (1 + 1/\sqrt{2})\eta TA_{\mathcal{X}}$$

minimized with $\eta = \sqrt{2\log(A_{\mathcal{X}})/(TA_{\mathcal{X}})}$. By summing everything, we now get, using this time the better inequality $H \leq \sqrt{A_{\mathcal{X}}}$ from Lemma E.2,

$$\mathfrak{R}_{\max}^T \leq (2 + 4\sqrt{2})\sqrt{TA_{\mathcal{X}}\iota} + (2 + 2\sqrt{2})\sqrt{TA_{\mathcal{X}}\log(A_{\mathcal{X}})}$$

and we conclude using $2\sqrt{2} \leq 3$ for readability. $\square$

F. Efficient updates and proof of correctness

Dilated entropy equivalence We first state a useful property related to the dilated divergence. For any list of learning rates $\eta^r := (\eta_h^r(x_h))_{h,x_h}$, we define the dilated entropy Ψ_{η^r} by

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

The following lemma is the straight-forward extension of the Lemma 9 by Kozuno et al. (2021) to non-constant learning rates.

33

Lemma F.1. *For any vector η^r of learning rates and $(\mu^1, \mu^2) \in \Pi^2_{\max}$, the Bregman divergence $\mathcal{D}_{\Psi_{\eta^r}}(\mu^1_{1:}, \mu^2_{1:})$ between the realization plans coincides with the dilated divergence $\mathcal{D}_{\eta^r}(\mu^1, \mu^2)$ defined in Section 5.*

The proof is essentially the same, and is based on the fact (later used in Proposition F.3) that for all positive $\mu \in \Pi_{\max}$

$$\nabla_{x_h, a_h} \Psi_{\eta^r}(\mu_{1:}) = \log(\mu_h(a_h | x_h)). \tag{6}$$

We then state the result on the equivalence between update (U1) and update (U2).

Proposition F.2. *Let $(\mu^t)_{t \in [|0, T|]}$ obtained using* `Balanced FTRL` *with Shannon entropy and any sequence $(\eta_h)_{h \in [H]}$ of learning rates. Then there exists a vector $\eta^\star$ of learning rates and policy $\mu^\star \in \Pi_{\max}$ that can be computed with a time complexity $\mathcal{O}(A_\mathcal{X})$ such that it coincides with update (U2),*

$$\mu^t = \arg\min_{\mu \in \Pi_{\max}} \left\langle \mu, \tilde{L}^{t-1} \right\rangle + D_{\eta^\star}(\mu, \mu^\star).$$

Proof. From update (U1), $\mu^t = \arg\min_{\mu \in \Pi_{\max}} \left\langle \mu_{1:}, \widetilde{L}^{t-1} \right\rangle + \Psi^1(\mu_{1:})$ where

$$\Psi^1(\mu_{1:}) = \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{p^\star_{1:h}(x_h) \mu_{1:h}(x_h, a_h)}{\eta_h} \log\left(p^\star_{1:h}(x_h) \mu_{1:h}(x_h, a_h)\right)$$

we first notice that

$$\Psi^1(\mu_{1:}) - \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{p^\star_{1:h}(x_h) \mu_{1:h}(x_h, a_h)}{\eta_h} \log\left(p^\star_{1:h}(x_h)\right)$$

$$= \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{p^\star_{1:h}(x_h) \mu_{1:h}(x_h, a_h)}{\eta_h} \log(\mu_{1:h}(x_h, a_h))$$

$$= \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \sum_{h'=1}^{h} \frac{p^\star_{1:h}(x_h) \mu_{1:h}(x_h, a_h)}{\eta_h} \log(\mu_{h'}(a_{h'} | x_{h'}))$$

$$= \sum_{h'=1}^{H} \sum_{(x_{h'}, a_{h'}) \in \mathcal{A}(\mathcal{X}_{h'})} \log(\mu_{h'}(a_{h'} | x_{h'})) \sum_{h=h'}^{H} \sum_{(\ldots, x_{h'}, a_{h'}, \ldots, x_h, a_h)} \frac{p^\star_{1:h}(x_h) \mu_{1:h}(x_h, a_h)}{\eta_h}$$

$$= \sum_{h'=1}^{H} \sum_{(x_{h'}, a_{h'}) \in \mathcal{A}(\mathcal{X}_{h'})} p^\star_{1:h'}(x_{h'}) \mu_{1:h'}(x_{h'}, a_{h'}) \log(\mu_{h'}(a_{h'} | x_{h'})) \sum_{h=h'}^{H} \frac{1}{\eta_h}$$

$$= \sum_{h'=1}^{H} \sum_{(x_{h'}, a_{h'}) \in \mathcal{A}(\mathcal{X}_{h'})} \frac{\mu_{1:h'}(x_{h'}, a_{h'})}{\eta^\star_{h'}} \log(\mu_{h'}(a_{h'} | x_{h'}))$$

$$= \Psi_{\eta^\star}(\mu_{1:})$$

where we defined

$$\eta^\star_{h'}(x_{h'}) := 1 / \left(p^\star_{1:h'}(x_{h'}) \sum_{h=h'}^{H} \frac{1}{\eta_h} \right).$$

This shows that Ψ^1 only differs from $\Psi_{\eta^\star}$ by a linear term of the realization plan: the Bregman divergence $\mathcal{D}_{\Psi^1}$ of Ψ^1 is thus exactly $\mathcal{D}_{\Psi_{\mu^\star}}$. We can then conclude using Lemma F.1 with $\mu^\star_{1:} := \arg\min_{\mu^\star \in \Pi^{1:}_{\max}} \Psi'(\mu^\star_{1:})$

$$\Psi^1(\mu_{1:}) = \mathcal{D}_{\Psi^1}(\mu_{1:}) + \Psi^1(\mu^\star_{1:}) = \mathcal{D}_{\Psi_{\mu^\star}}(\mu, \mu^\star) + \Psi'(\mu^\star_{1:}) = \mathcal{D}_{\mu^\star}(\mu, \mu^\star) + \Psi'(\mu^\star_{1:}).$$

Algorithm 3 Fast Tree Update

1: **Input:**
μ^{t-1} given by update (U2)
Trajectory $(x_1^{t-1}, a_1^{t-1}, ..., x_H^{t-1}, a_H^{t-1})$
Loss $\widetilde{\ell}^{t-1}$, old and new learning rates η^{t-1} and η^t along the trajectory
Base policy μ^0

2: **Init:**
$Z_{H+1}^{t-1} \leftarrow 1$
$\alpha_h^{t-1} = \eta_h^t(x_h^{t-1})/\eta_h^{t-1}(x_h^{t-1})$

3: **Iterate:**
For $h = H$ to 1
$\overline{\ell}_h^{t-1} \leftarrow \widetilde{\ell}_h^{t-1}(x_h^{t-1}, a_h^{t-1}) - \log(Z_{h+1}^{t-1})/\eta_{h+1}^t(x_{h+1}^{t-1})$
$Z_h^{t-1} \leftarrow \sum_{a_h \in \mathcal{A}(x_h^t)} \mu_h^t(a_h|x_h^{t-1})^{\alpha_h^{t-1}} \mu_h^0(a_h|x_h^{t-1})^{1-\alpha_h^{t-1}} \exp\left(-\mathbb{I}_{\{a_h = a_h^{t-1}\}} \eta_h^t(x_h^{t-1})\overline{\ell}_h^t\right)$
For all $a_h \in \mathcal{A}\left(x_h^{t-1}\right)$:
$\mu_h^t(a_h|x_h^{t-1}) \leftarrow \mu_h^{t-1}(a_h|x_h^{t-1})^{\alpha_h^{t-1}} \mu_h^0(a_h|x_h^{t-1})^{1-\alpha_h^{t-1}} \exp\left(-\log(Z_h^{t-1}) - \mathbb{I}_{\{a_h = a_h^{t-1}\}} \eta_h^t(x_h^{t-1})\overline{\ell}_h^t\right)$
For all other information sets x_h :
$\mu_h^t(.|x_h) \leftarrow \mu_h^{t-1}(.|x_h)$

4: **Output:** μ^t

This shows that the two updates minimize two functions of μ that only differ by a constant terms, and are thus equivalent.

The calculation of the $\eta^\star$ is straight-forward with the above formula, while the policy $\mu^\star$ is the minimum over a dilated entropy function that can be recursively computed in $\mathcal{O}(A_\mathcal{X})$, starting from the leaves. $\square$

We now state Algorithm 3 and a proof of its correctness.

Proposition F.3. *Assume that for all $t \in [T]$, the vector η^t of learning rates is only updated along the trajectory of episode t. Let $(\mu^t)_{t \in [T]}$ be the sequence of policies computed with update (U2). Then it coincides with the sequence of policies that would be computed with Algorithm 3.*

Proof. We remind that update (U2) define the sequence $(\mu^t)_{t \in [T]}$ with

$$\mu^t = \operatorname{argmin}_{\mu \in \Pi_{\max}} \left\langle \mu_{1:}, \widetilde{L}^{t-1} \right\rangle + \mathcal{D}_{\eta^t}(\mu, \mu^0).$$

Using the decomposition $\Pi_{\max}^{1:} = (F + u) \cap \mathbb{R}_{\geq 0}^{A_\mathcal{X}}$, Lemma F.1 (on the equivalence between $\mathcal{D}_{\eta^t}$ and $\mathcal{D}_{\Psi_{\eta^t}}$), we then know the existence for all $t \in [T]$ of $g^t \in F^\perp$ such that

$$\nabla\Psi_{\eta^t}(\mu^t) - \nabla\Psi_{\eta^t}(\mu^0) - \widetilde{L}^{t-1} = g^t.$$

Taking the difference on the last equation between t and $t - 1$ yields:

$$\nabla\Psi_{\eta^t}(\mu^t) = \nabla\Psi_{\eta^{t-1}}(\mu^{t-1}) + \Psi_{\eta^{t-1}}(\mu^0) - \Psi_{\eta^t}(\mu^0) - \widetilde{\ell}^{t-1} + \Delta_t \tag{7}$$

where $\Delta_t := g_t - g_{t-1} \in F^\perp$. And from the normalizing constraints on $\Pi_{\max}^{1:}$, Δ_t is characterized by the vector $\lambda^t = (\lambda^t(x))_{x \in \mathcal{X}}$ with

$$\Delta_t(x_h, a_h) = -\lambda^t(x_h) + \sum_{(x_{h+1}) \in \mathcal{X}_{h+1}(x_h, a_h)} \lambda(x_{h+1})$$

with the sum on the second term being null when $h = H$.

We first show by backward induction on $h \in [H]$ that for all x_h not visited at episode $t - 1$, $\lambda^t(x_h) = 0$. Indeed, in this case $\eta_h^{t-1}(x_h) = \eta_h^t(x_h)$, and the induction property leads to $\Delta^t(x_h, a_h) = -\lambda^t(x_h)$ for all a_h, as the perfect recall assumption

35

implies any x_{h+1} is also not visited at episode t. For all $a_h \in \mathcal{A}(x_h)$, using equation (7) on the component (x_h, a_h) gives with the expression of the gradient of equation (6)

$$\frac{1}{\eta_h^t(x_h)} \log(\mu_h^t(a_h|x_h)) = \frac{1}{\eta_h^t(x_h)} \log(\mu_h^{t-1}(a_h|x_h)) - \lambda^t(x_h)$$

which is equivalent to

$$\mu_h^t(a_h|x_h)) = \mu_h^{t-1}(a_h|x_h) \exp\left(-\eta_h^t(x_h)\lambda_h^t(x_h)\right).$$

Summing over $a_h \in \mathcal{A}(x_h)$ yields $\eta_h^t(x_h)\lambda_h^t(x_h) = 0$ and thus $\lambda_h^t(x_h) = 0$ as $\eta_h^t(x_h) \neq 0$, which concludes the induction. We finally get from the previous equality $\mu_h^t(.|x_h) = \mu_h^{t-1}(.|x_h)$ for all x_h outside of the trajectory at time $t-1$, as in Algorithm 3.

For all $h \in [H]$, we are then interested in the visited state x_h^{t-1}, and define by convention $\lambda^t(x_{H+1}^t) = 0$. Then we have for $a_h \in \mathcal{A}\left(x_h^{t-1}\right)/\{a_h^{t-1}\}$, $\Delta^t(x_h^{t-1}, a_h) = -\lambda^t(x_h^{t-1})$ and $\Delta^t(x_h^{t-1}, a_h^{t-1}) = -\lambda^t(x_h^{t-1}) + \lambda^t(x_{h+1}^{t-1})$ using the previous property that λ^t is null on unvisited states. For all $a_h \in \mathcal{A}\left(x_h^{t-1}\right)$, we now have with equations (6) and (7):

$$\frac{1}{\eta_h^t(x_h^{t-1})} \log(\mu_h^t(a_h|x_h^{t-1})) = \frac{1}{\eta_h^{t-1}(x_h^{t-1})} \log(\mu_h^{t-1}(a_h|x_h^{t-1})) + \left(\frac{1}{\eta_h^t(x_h^{t-1})} - \frac{1}{\eta_h^{t-1}(x_h^{t-1})}\right) \log(\mu_h^0(a_h|x_h^{t-1}))$$

$$+ \mathbb{I}_{\{a_h = a_h^{t-1}\}} \left(\lambda^t(x_{h+1}^{t-1}) - \widetilde{\ell}_h^{t-1}(x_h^{t-1}, a_h^{t-1})\right) - \lambda^t(x_h^{t-1}).$$

Multiplying by $\eta_h^t(x_h^{t-1})$ and taking the exponential then leads to:

$$\mu_h^t(a_h|x_h^{t-1}) = \mu_h^{t-1}(a_h|x_h^{t-1})^{\alpha_h^{t-1}} \mu_h^0(a_h|x_h^{t-1})^{1-\alpha_h^{t-1}} \exp\left(-\eta_h^t(x_h^{t-1})\lambda^t(x_h^{t-1}) - \mathbb{I}_{\{a_h = a_h^{t-1}\}} \eta_h^t(x_h^{t-1})\overline{\ell}_h^t\right)$$

where $\overline{\ell}_h^{t-1} = \widetilde{\ell}_h^{t-1}(x_h^{t-1}, a_h^{t-1}) - \lambda^t(x_{h+1}^{t-1})$ and $\alpha_h^{t-1} = \eta_h^t(x_h^{t-1})/\eta_h^{t-1}(x_h^{t-1})$. We again recognize the update of algorithm 3 with $Z_h^{t-1} = \exp\left(\eta_h^t(x_h^{t-1})\lambda^t(x_h^{t-1})\right)$, as summing the previous equation over $a_h \in \mathcal{A}\left(x_h^{t-1}\right)$ yields:

$$Z_h^{t-1} = \sum_{a_h \in \mathcal{A}\left(x_h^{t-1}\right)} \mu_h^t(a_h|x_h^{t-1})^{\alpha_h^{t-1}} \mu_h^0(a_h|x_h^{t-1})^{1-\alpha_h^{t-1}} \exp\left(-\mathbb{I}_{\{a_h = a_h^{t-1}\}} \eta_h^t(x_h^{t-1})\overline{\ell}_h^t\right).$$

We therefore obtained that policy μ^t obtained with update (U2) coincides with the output of Algorithm 3 both in visited and unvisited information set at episode $t-1$, which concludes the proof. $\qquad \square$

G. Proof of `Adaptive FTRL`

In all this section, we will use the constants $\iota' = \log(3A_{\mathcal{X}}T/\delta)$ and $\upsilon = 1 + \log_2(T)$ of theorem 5.1 for readability.

This first lemma centralizes the necessary inequalities for the analysis of the regret which can all hold with a high probability.

Lemma G.1. *We assume $\iota'/\gamma \leq 1/2$ and $\gamma \geq 1$. Then, when using* `Adaptive FTRL`*, the following inequalities all hold with a probability at least $1 - \delta$:*

$$\widetilde{P}_{1:}^T(x_h, a_h) \leq 1 + 2P_{1:}^T(x_h), \tag{8}$$

$$\widetilde{L}^T(x_h, a_h) - L^T(x_h, a_h) \leq 1 + \frac{2\iota'}{\gamma} P_{1:}^T(x_h), \tag{9}$$

$$\sum_{t=1}^T \left\langle \mu_{1:}^t, \ell^t - \widetilde{\ell}^t \right\rangle \leq H\sqrt{2T\iota'}. \tag{10}$$

We denote by $\mathcal{E}$ the event under which all these inequalities hold.

Proof. Let $(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)$. We can first notice that for all $0 \leq t \leq T$, $1 + \widetilde{P}_{1:h}^t(x_h, a_h) \leq 2^t$. Indeed, $1 + \widetilde{P}_a^0 = 1$ and recursively, if $1 + \widetilde{P}_{1:h}^t(x_h, a_h) \leq 2^t$, then $\widetilde{p}_{1:h}^{t+1}(x_h, a_h) \leq (1 + \widetilde{P}_{1:h}^t(x_h, a_h))/\gamma \leq 2^t/\gamma \leq 2^t$, which implies $1 + \widetilde{P}_{1:h}^{t+1}(x_h, a_h) \leq 2^{t+1}$.

In particular, we get $\gamma_h^T(x_h, a_h) \geq \gamma/2^T$, and define for all $i \in [T]$, $\gamma_i^\triangledown = \gamma/2^i$.

Let $\tau_h^i(x_h, a_h) = \max\{t \leq T, \gamma_h^t(x_h, a_h) \geq \gamma_i^\triangledown\}$. As $\gamma_h^t(x_h, a_h)$ is a decreasing predictable sequence for $\mathcal{F}_t$, $\tau_h^i(x_h, a_h)$ is a stopping time, which lets us apply Lemma D.1 with $\tau_h(x_h, a_h) = \tau_h^i(x_h, a_h)$, $\delta' = \delta/(3A_\mathcal{X}T)$ and $\gamma_h'(x_h, a_h) = \gamma_i^\triangledown$ to get with a probability at least $1 - \delta'$,

$$\widetilde{L}_h^{\tau_h^i(x_h, a_h)}(x_h, a_h) - L_h^{\tau_h^i(x_h, a_h)}(x_h, a_h) \leq \frac{\iota'}{2\gamma_i^\triangledown}$$

where $\iota' = \log(3A_\mathcal{X}T/\delta)$. We especially get that this inequality holds for all $(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)$ and $i \in [T]$ with a probability at least $1 - \delta/3$. As $\gamma/2^T \leq \gamma_h^T(x_h, a_h) \leq \gamma$, we know the existence of $i(x_h, a_h)$, such that $\gamma_{i(x_h, a_h)}^\triangledown \leq \gamma_h^T(x_h, a_h) \leq 2\gamma_{i(x_h, a_h)}^\triangledown$, and the previous inequality used with $i(x_h, a_h)$ then gives:

$$\widetilde{L}_h^T(x_h, a_h) - L_h^T(x_h, a_h) \leq \frac{\iota'}{2\gamma_i^\triangledown} \leq \frac{\iota'}{\gamma_h^T(x_h, a_h)} \; .$$

As the definition of $i(x_h, a_h)$ implies $\tau_h^{i(x_h, a_h)}(x_h, a_h) = T$

The same exact steps (along with Lemma D.1) can also be performed with $\widetilde{p}_{1:h}^t(x_h, a_h)$, and we similarly obtain that with probability at least $1 - \delta/3$, for all $(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)$,

$$\widetilde{P}_{1:h}^T(x_h, a_h) - P_{1:h}^T(x_h, a_h) \leq \frac{\iota'}{\gamma_h^T(x_h, a_h)}$$

which gives, using the definition of $\gamma_h^T(x_h, a_h)$,

$$\widetilde{P}_{1:h}^T(x_h, a_h) - P_{1:h}^T(x_h, a_h) \leq \frac{\iota'}{\gamma}\left(1 + \widetilde{P}_{1:h}^{T-1}(x_h, a_h)\right) \leq \frac{\iota'}{\gamma}\left(1 + \widetilde{P}_{1:h}^T(x_h, a_h)\right)$$

and, as $\frac{\iota'}{\gamma} \leq 1/2$ by assumption,

$$\widetilde{P}_{1:h}^T(x_h, a_h) \leq \frac{2\iota'}{\gamma} + 2P_{1:h}^T(x_h, a_h) \leq 1 + 2P_{1:h}^T(x_h, a_h)$$

and we recognize inequality (8). Plugging this back in the previous inequalities on $\widetilde{L}_h^T$ also yields

$$\widetilde{L}_h^T(x_h, a_h) - L_h^T(x_h, a_h) \leq \frac{\iota'}{\gamma}\left(1 + \widetilde{P}_{1:h}^T(x_h, a_h)\right) \leq \frac{\iota'}{\gamma}\left(2 + 2P_{1:h}^T(x_h, a_h)\right) \leq 1 + \frac{2\iota'}{\gamma}P_{1:h}^T(x_h, a_h)$$

and we now recognize inequality (9).

The last inequality is a consequence of the Lemma D.2 used on $u_t = \left\langle \mu_{1:}^t, \widehat{\ell}^t \right\rangle$. Indeed, as u_t is non-negative and $\mathbb{E}\left[\left\langle \mu_{1:}^t, \widehat{\ell}^t \right\rangle | \mathcal{F}_{t-1}\right] = \langle \mu_{1:}^t, \ell^t \rangle \leq H$, the lemma gives with a probability at least $1 - \delta/3$

$$\sum_{t=1}^T \left\langle \mu_{1:}^t, \ell^t - \widetilde{\ell}^t \right\rangle \leq H\sqrt{2T}\log(3/\delta) \leq H\sqrt{2T\iota'} \; .$$

As the three inequalities each hold with a probability at least $1 - \delta/3$, they all hold at the same time with a probability at least $1 - \delta$, which concludes the proof. $\qquad\square$

As in Section E, we decompose the regret between the BIAS, REG and VAR terms. Note that the analysis of REG and VAR terms are different, although conceptually the same. The proof is directly based on the algorithmic update (3), in which the sequences $(Z_h^t)_{h \in [H]}$ are defined.

Lemma G.2. *Using* `Adaptive FTRL` *yields*

$$\mathfrak{R}_{\max}^T \leq \underbrace{\sum_{t=1}^{T} \left\langle \mu_{1:}^t, \ell^t - \widetilde{\ell}^t \right\rangle}_{BIAS\ I} + \underbrace{\max_{\mu_{1:}^\dagger \in \Pi_{\max}} \sum_{t=1}^{T} \left\langle \mu^\dagger, \widetilde{\ell}^t - \ell^t \right\rangle}_{BIAS\ II}$$

$$+ \underbrace{\max_{\mu^\dagger \in \Pi_{\max}} \mathcal{D}_{\eta^{T+1}} \left(\mu^\dagger, \mu^0 \right)}_{REG} + \underbrace{\sum_{t=1}^{T} \left[\left\langle \mu_{1:}^t, \widetilde{\ell}^t \right\rangle + \frac{1}{\eta_1^{t+1}(x_1^t)} \log \left(Z_1^t \right) \right]}_{VAR}.$$

Proof. Let $t \in [T]$ and $h \in [H]$. From Algorithm 3, we have for all $a_h \in \mathcal{A}(x_h^t)$,

$$\mu_h^{t+1}(x_h^t, a_h) = \mu_h^t(a_h|x_h^t)^{\alpha_h^t} \mu_h^0(a_h|x_h^t)^{1-\alpha_h^t} \exp\left(-\log(Z_h^t) - \mathbb{I}_{\{a_h = a_h^t\}} \eta_h^{t+1}(x_h^t) \overline{\ell}_h^t \right)$$

where $\overline{\ell}_h^t = \widetilde{\ell}_h^t - \frac{\log Z_{h+1}^t}{\overline{\eta}_{h+1}^t}$. We then get, for all $\mu^\dagger \in \Pi_{\max}$,

$$\sum_{a_h \in \mathcal{A}(x_h^t)} \mu_{1:h}^\dagger(x_h^t, a_h) \log(\mu_h^{t+1}(a_h|x_h^t)) = \sum_{a_h \in \mathcal{A}(x_h^t)} \mu_{1:h}^\dagger(x_h, a_h) \left[\alpha_h^t \log(\mu_h^t(a_h|x_h^t)) + (1-\alpha_h^t) \log(\mu_h^0(a_h|x_h^t)) - \log(Z_h^t) \right]$$

$$- \eta_h^{t+1}(x_h^t) \left[\mu_{1:h}^\dagger(x_h^t, a_h^t) \widetilde{\ell}_h^t + \frac{\eta_h^{t+1}(x_h^t)}{\mu_{1:h}^\dagger(x_h^t, a_h)} \log(Z_{h+1}^t) \right].$$

Using this equality then gives, with the convention $\mu_{1:0}^\dagger(.) = 1$,

$$-\left\langle \mu_{1:}^\dagger, \widetilde{\ell}^t \right\rangle = -\sum_{h=1}^{H} \mu_{1:h}^\dagger(x_h^t, a_h^t)^\dagger \widetilde{\ell}_h^t$$

$$= \sum_{h=1}^{H} \sum_{a_h \in \mathcal{A}(x_h^t)} \frac{\mu_{1:h}^\dagger(x_h, a_h)}{\eta_h^{t+1}(x_h^t)} \left[\log(\mu_h^{t+1}(a_h|x_h^t)) - \alpha_h^t \log(\mu_h^t(a_h|x_h^t)) - (1-\alpha_h^t) \log(\mu_h^0(a_h|x_h^t)) + \log(Z_h^t) \right]$$

$$- \sum_{h=1}^{H} \frac{\mu_{1:h}^\dagger(a_h^t|x_h^t)}{\eta_{h+1}^{t+1}} \log(Z_{h+1}^t)$$

$$= \sum_{h=1}^{H} \sum_{a_h \in \mathcal{A}(x_h^t)} \frac{\mu_{1:h}^\dagger(x_h, a_h)}{\eta_h^{t+1}(x_h^t)} \left[-\log\left(\frac{\mu_h^\dagger(a_h|x_h^t)}{\mu_h^{t+1}(a_h|x_h^t)} \right) + \alpha_h^t \log\left(\frac{\mu_h^\dagger(a_h|x_h^t)}{\mu_h^t(a_h|x_h^t)} \right) + (1-\alpha_h^t) \log\left(\frac{\mu_h^\dagger(a_h|x_h^t)}{\mu_h^0(a_h|x_h^t)} \right) \right]$$

$$+ \sum_{h=1}^{H} \left[\frac{\mu_{1:h-1}^\dagger(a_{h-1}^t|x_{h-1}^t)}{\eta_h^{t+1}(x_h^t)} \log(Z_h^t) - \frac{\mu_{1:h}^\dagger(a_h^t|x_h^t)}{\eta_{h+1}^{t+1}(x_{h+1}^t)} \log(Z_{h+1}^t) \right]$$

$$= \sum_{h=1}^{H} \sum_{a_h \in \mathcal{A}(x_h^t)} \mu_{1:h}^\dagger(x_h, a_h) \left[\frac{-1}{\eta_h^{t+1}(x_h^t)} \log\left(\frac{\mu_h^\dagger(a_h|x_h^t)}{\mu_h^{t+1}(a_h|x_h^t)} \right) + \frac{1}{\eta_h^t} \log\left(\frac{\mu_h^\dagger(a_h|x_h^t)}{\mu_h^t(a_h|x_h^t)} \right) \right.$$

$$\left. + \left(\frac{1}{\eta_h^{t+1}(x_h^t)} - \frac{1}{\eta_h^t(x_h^t)} \right) \log\left(\frac{\mu_h^\dagger(a_h|x_h^t)}{\mu_h^0(a_h|x_h^t)} \right) \right] + \frac{1}{\eta_1^{t+1}(x_1^t)} \log(Z_1^t).$$

Then, using that $\eta_h^t(x_h)$ only updates when $x_h = x_h^t$ and in the second equality telescoping, with $\mu^1 = \mu^0$, we get

38

$$-\sum_{t=1}^{T}\left\langle \mu_{1:}^{\dagger},\widetilde{\ell}^{t}\right\rangle = \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\mu_{1:h}^{\dagger}(x_h,a_h)\left[-\frac{-1}{\eta_h^{t+1}(x_h)}\log\left(\frac{\mu_h^{\dagger}(a_h|x_h)}{\mu_h^{t+1}(a_h|x_h)}\right) + \frac{1}{\eta_h^{t}(x_h)}\log\left(\frac{\mu_h^{\dagger}(a_h|x_h)}{\mu_h^{t}(a_h|x_h)}\right)\right.$$

$$\left. + \left(\frac{1}{\eta_h^{t+1}(x_h)} - \frac{1}{\eta_h^{t}}(x_h)\right)\log\left(\frac{\mu_a^{\dagger}}{\mu_h^{0}(x_h^t,a_h)}\right)\right] + \sum_{t=1}^{T}\frac{1}{\eta_1^{t+1}(x_1^t)}\log(Z_1^t)$$

$$= -\mathcal{D}_{\eta^{T+1}}\left(\mu^{\dagger},\mu^{T+1}\right) + \mathcal{D}_{\eta^{T+1}}\left(\mu^{\dagger},\mu^{0}\right) + \sum_{t=1}^{T}\frac{1}{\eta_1^{t+1}(x_1^t)}\log(Z_1^t)$$

$$\leq \mathcal{D}_{\eta^{T+1}}\left(\mu^{\dagger},\mu^{0}\right) + \sum_{t=1}^{T}\frac{1}{\eta_1^{t+1}(x_1^t)}\log(Z_1^t)\,.$$

The rest of the lemma then follows by the decomposition

$$\left\langle \mu_{1:}^{t} - \mu_{1:}^{\dagger},\ell^{t}\right\rangle = \left\langle \mu_{1:}^{t},\ell^{t} - \widetilde{\ell}^{t}\right\rangle + \left\langle \mu_{1:}^{\dagger},\widetilde{\ell}^{t} - l^{t}\right\rangle + \left\langle \mu_{1:}^{t} - \mu^{\dagger},\widetilde{\ell}^{t}\right\rangle$$

and the expression of the regret as a scalar product. $\qquad\square$

The following lemma gives a bound with high probability on the two BIAS terms for `Adaptive FTRL`.

Lemma G.3. *Under event $\mathcal{E}$, assuming $\gamma \geq 1$, we have*

$$BIAS\ I \leq H\sqrt{2T\iota'} + \gamma\upsilon A_{\mathcal{X}}\,, \qquad BIAS\ II \leq X + \frac{2\iota'}{\gamma}HT\,.$$

Proof. We first decompose BIAS I into two terms,

$$\sum_{t=1}^{T}\left\langle \mu_{1:}^{t},\ell^{t} - \widetilde{\ell}^{t}\right\rangle = \sum_{t=1}^{T}\left\langle \mu_{1:}^{t},\ell^{t} - \widehat{\ell}^{t}\right\rangle + \sum_{t=1}^{T}\left\langle \mu_{1:}^{t},\widehat{\ell}^{t} - \widetilde{\ell}^{t}\right\rangle\,.$$

The first term is directly upper-bounded by $H\sqrt{2T\iota'}$ using the inequality (10). For the second term, we have

$$\sum_{t=1}^{T}\left\langle \mu_{1:}^{t},\widehat{\ell}^{t} - \widetilde{\ell}^{t}\right\rangle = \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\mathbb{I}_{\{x_h=x_h^t,a_h=a_h^t\}}\mu_{1:h}^{t}(x_h,a_h)\left(\frac{1-r_h^t}{\mu_{1:h}^{t}(x_h,a_h)} - \frac{1-r_h^t}{\mu_{1:h}^{t}(x_h,a_h) + \gamma_h^t(x_h,a_h)}\right)$$

$$= \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\mathbb{I}_{\{x_h=x_h^t,a_h=a_h^t\}}(1-r_h^t)\frac{\gamma_h^t(x_h,a_h)}{\mu_{1:h}^{t}(x_h,a_h) + \gamma_h^t(x_h,a_h)}$$

$$\leq \sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\widetilde{p}_{1:h}^{t}(x_h,a_h)\gamma_h^t(x_h,a_h)$$

$$= \gamma\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\sum_{t=1}^{T}\frac{\widetilde{p}_{1:h}^{t}(x_h,a_h)}{1 + \widetilde{P}_{1:h}^{t-1}(x_h,a_h)}$$

$$\leq_1 \gamma\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\log_2(1 + \widetilde{P}_{1:h}^{T}(x_h,a_h))$$

$$\leq_2 \gamma\sum_{h=1}^{H}\sum_{(x_h,a_h)\in\mathcal{A}(\mathcal{X}_h)}\log_2(2 + 2P_{1:h}^{T}(x_h,a_h))$$

$$\leq \gamma\upsilon A_{\mathcal{X}}$$

as $\upsilon = 1 + \log_2(1 + T)$ by definition, where we used Lemma D.3 for $\leq_1$, as $\widetilde{p}_{1:h}^t(x_h, a_h) \leq \frac{1}{\gamma}(1 + \widetilde{P}_{1:h}^{t-1}(x_h, a_h)) \leq 1 + \widetilde{P}_{1:h}^{t-1}(x_h, a_h)$ and inequality (8) for $\leq_2$.

In order to bound BIAS II, we use inequality (9) and get for all $\mu^\dagger \in \Pi_{\max}$,

$$\left\langle \mu_{1:,}^\dagger, \widetilde{L}^T - L^T \right\rangle \leq \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \mu_{1:h}^\dagger(x_h, a_h) + \frac{2\iota'}{\gamma} \left\langle \mu_{1:,}^\dagger, P_{1:}^T \right\rangle \leq X + \frac{2\iota'}{\gamma} HT \,.$$

$\square$

Contrary to the analysis of the first algorithm, the bound on the REG term now only holds with high probability because of the adaptive learning rate.

Lemma G.4. *Under the event $\mathcal{E}$, we have the bound*

$$REG \leq \frac{2H \log(A_\mathcal{X})}{\eta}(X + T) \,.$$

Proof. Using inequality (8) and the fact that μ^0 is the uniform distribution, we get for any $\mu^\dagger \in \Pi_{\max}$

$$\mathcal{D}_{\eta^{T+1}}\left(\mu^\dagger, \mu^0\right) = \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \frac{1}{\eta_h^{T+1}(x_h)} \mu_{1:h}^\dagger(x_h, a_h) \log\left(\frac{\mu_h^\dagger(a_h|x_h)}{\mu_h^0(a_h|x_h)}\right)$$

$$\leq \frac{1}{\eta} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} \max_{x_{h'}' \geq x_h} \left[1 + \widetilde{P}_{1:h'}^T(x_{h'}')\right] \sum_{a_h \in \mathcal{A}(x)} \mu_{1:h}^\dagger(x_h, a_h) \log(A_x)$$

$$\leq \frac{\log(A_\mathcal{X})}{\eta} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} \max_{x_{h'}' \geq x_h} \left[1 + \frac{1}{A_x} \sum_{a_{h'}' \in \mathcal{A}x_{h'}} \left(1 + \widetilde{P}_{1:h'}^T(x_{h'}', a_{h'}')\right)\right] \sum_{a_h \in \mathcal{A}(x)} \mu_{1:h}^\dagger(x_h, a_h)$$

$$\leq \frac{\log(A_\mathcal{X})}{\eta} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} \max_{x_{h'}' \geq x_h} \left[2 + 2P_{1:h'}^T(x_{h'}')\right] \sum_{a_h \in \mathcal{A}(x)} \mu_{1:h}^\dagger(x_h, a_h)$$

$$\leq \frac{2\log(A_\mathcal{X})}{\eta} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} \left[1 + P_{1:h}^T(x_h)\right] \sum_{a_h \in \mathcal{A}(x)} \mu_{1:h}^\dagger(x_h, a_h)$$

$$\leq \frac{2H \log(A_\mathcal{X})}{\eta}(X + T)$$

and we concludes taking the max over $\mu^\dagger \in \Pi_{\max}$. $\square$

Upper bounding the VAR term is however more difficult. We give a preliminary lemma that requires the assumption that the individual learning rates of the information states at any given time increase along a trajectory. This hypothesis is trivially satisfied by `Adaptive FTRL` thanks to the particular definition of η_h^t.

Lemma G.5. *For all $t \in [T]$, if the learning rates at time t increase along the trajectory, then*

$$\left\langle \mu_{1:,}^t, \widetilde{\ell}^t \right\rangle + \frac{1}{\eta_1^{t+1}(x_1^t)} \log\left(Z_1^t\right) \leq \frac{H}{2} \sum_{h=1}^{H} \eta_h^t(x_h^t) \widetilde{\ell}_h^t(x_h^t, a_h^t) \,.$$

Proof. For readability we will define for all $h \in [H]$ in this proof, $\overline{\mu}_h^t := \mu_h^t(a_h^t|x_h^t), \overline{\eta}_h^t := \eta_h^t(x_h^t), \kappa_h^t := \overline{\eta}_h^t/\overline{\eta}_{h+1}^t$ and $\widetilde{\ell}_h^t := \widetilde{\ell}_h^t(x_h^t, a_h^t)$.

Using Hölder inequality, we first get:

40

$$Z_h^t = \sum_{a_h \in \mathcal{A}\left(x_h^t\right)} \mu_h^t(a_h|x_h^t)^{\alpha_h^t} \mu_h^0(a_h|x_h^t)^{1-\alpha_h^t} \exp\left[-\mathbb{I}_{\left\{a_h^t = a_h\right\}} \eta_h^{t+1}(x_h^t)\overline{\ell}_h^t\right]$$

$$\leq \left(\sum_{a_h \in \mathcal{A}\left(x_h^t\right)} \mu_h^t(a_h|x_h^t)) \exp\left[-\mathbb{I}_{\left\{a_h = a_h^t\right\}} \eta_h^{t+1}(x_h^t)\overline{\ell}_h^t/\alpha_h^t\right]\right)^{\alpha_h^t} \left(\sum_{a_h \in \mathcal{A}\left(x_h^t\right)} \mu_h^0(a_h|x_h^t))\right)^{1-\alpha_h^t}$$

$$= \left(1 - \overline{\mu}_h^t + \overline{\mu}_h^t \exp\left[-\eta_h^{t+1}(x_h^t)\overline{\ell}_h^t/\alpha_h^t\right]\right)^{\alpha_h^t}.$$

This yields as $\overline{\ell}_h^t = \widetilde{\ell}_h^t - \log(Z_{h+1}^t)/\eta_{h+1}^{t+1}(x_{h+1}^t)$ and $\alpha_h^t = \eta_h^{t+1}(x_h^t)/\overline{\eta}_h^t$, using convexity and $\kappa_h^t \leq 1$,

$$(Z_h^t)^{1/\alpha_h^t} \leq 1 - \overline{\mu}_h^t + \overline{\mu}_h^t \left(Z_{h+1}^t\right)^{\overline{\eta}_h^t/\eta_{h+1}^{t+1}(x_{h+1}^t)} \exp\left(-\overline{\eta}_h^t \widetilde{\ell}_h^t\right)$$

$$= 1 - \overline{\mu}_h^t + \overline{\mu}_h^t \left(Z_{h+1}^t\right)^{\kappa_h^t/\alpha_{h+1}^t} \exp\left(-\overline{\eta}_h^t \widetilde{\ell}_h^t\right)$$

$$\leq 1 - \overline{\mu}_h^t + \overline{\mu}_h^t \left(\kappa_h^t \left(Z_{h+1}^t\right)^{1/\alpha_{h+1}^t} + (1 - \kappa_h^t)\right) \exp\left[-\overline{\eta}_h^t \widetilde{\ell}_h^t\right].$$

We can then define the sequence W_h^t such that $W_{H+1}^t := 1$, and for which the previous relation is always an equality, i.e.

$$W_h^t := 1 - \overline{\mu}_h^t + \overline{\mu}_h^t \left(\kappa_h^t W_{h+1}^t + (1 - \kappa_h^t)\right) \exp\left[-\overline{\eta}_h^t \widetilde{\ell}_h^t\right]$$

and we can get by induction, for all $h \in [H]$, $W_h^t \geq (Z_h^t)^{1/\alpha_h^t}$. If we introduce the independent auxiliary variables $y_h^t \sim \mathcal{B}(\kappa_h^t)$ and $z_h^t \sim \mathcal{B}(\overline{\mu}_h^t)$, with the associated products $y_{h:h'}^t = \Pi_{h \leq i \leq h'} y_i^t$ and $z_{h:h'}^t = \Pi_{h \leq i \leq h'} z_i^t$, we recursively get

$$W_h^t = \mathbb{E}_{y^t, z^t} \left[\exp\left(-\sum_{h \leq h' \leq H} y_{h:h'-1}^t z_{h:h'}^t \overline{\eta}_{h'}^t \widetilde{\ell}_{h'}^t\right)\right].$$

Finally, $\log(x) \leq -1 + x$, $e^{-x} \leq 1 - x + x^2/2$ and the above relation leads to

$$\log(Z_1^t) \leq \alpha_1^t \log(W_1^t)$$

$$= \alpha_1^t \log\left(\mathbb{E}_{y^t, z^t}\left[\exp\left(-\sum_{h=1}^{H} y_{1:h-1}^t z_{1:h}^t \overline{\eta}_h^t \widetilde{\ell}_h^t\right)\right]\right)$$

$$\leq \alpha_1^t \left(-1 + \mathbb{E}_{y^t, z^t}\left[\exp\left(-\sum_{h=1}^{H} y_{1:h-1}^t z_{1:h}^t \overline{\eta}_h^t \widetilde{\ell}_h^t\right)\right]\right)$$

$$\leq -\alpha_1^t \mathbb{E}_{y^t, z^t}\left[\sum_{h=1}^{H} y_{1:h-1}^t z_{1:h}^t \overline{\eta}_h^t \widetilde{\ell}_h^t\right] + \frac{1}{2}\alpha_1^t \mathbb{E}_{y^t, z^t}\left[\left(\sum_{h=1}^{H} y_{1:h-1}^t z_{1:h}^t \overline{\eta}_h^t \widetilde{\ell}_h^t\right)^2\right]$$

$$\leq -\alpha_1^t \sum_{h=1}^{H} \overline{\eta}_1^t \overline{\mu}_{1:h}^t \widetilde{\ell}_h^t + \frac{1}{2}\alpha_1^t \mathbb{E}_{y^t, z^t}\left[H \sum_{h=1}^{H} y_{1:h-1}^t z_{1:h}^t \left(\overline{\eta}_h^t \widetilde{\ell}_h^t\right)^2\right]$$

$$= -\eta_1^{t+1}(x_1^t) \sum_{h=1}^{H} \overline{\mu}_{1:h}^t \widetilde{\ell}_h^t + \frac{H}{2}\alpha_1^t \sum_{h=1}^{H} \overline{\eta}_1^t \overline{\mu}_{1:h}^t \overline{\eta}_h^t \left(\widetilde{\ell}_h^t\right)^2$$

$$\leq -\eta_1^{t+1}(x_1^t) \sum_{h=1}^{H} \overline{\mu}_{1:h}^t \widetilde{\ell}_h^t + \frac{H}{2}\eta_1^{t+1}(x_1^t) \sum_{h=1}^{H} \overline{\eta}_h^t \widetilde{\ell}_h^t$$

where we used $\widetilde{\ell}_h^t \overline{\mu}_{1:h}^t \leq 1$ for the last inequality, and we finally obtain

$$\sum_{h=1}^{H} \overline{\mu}_{1:h}^t \widetilde{\ell}_h^t + \frac{1}{\eta_1^{t+1}(x_1^t)} \log(Z_1^t) \leq \frac{H}{2} \sum_{h=1}^{H} \overline{\eta}_h^t \widetilde{\ell}_h^t$$

which concludes the proof. □

We then directly use this lemma on all episodes to upper bound the VAR term.

Lemma G.6. *Under the event $\mathcal{E}$, assuming $\gamma \geq 1$, we have the bound*

$$VAR \leq \frac{\eta}{2} H A_{\mathcal{X}} \upsilon .$$

Proof. Thanks to Lemma G.5, we have

$$\begin{aligned}
VAR &\leq \frac{H}{2} \sum_{t=1}^{T} \sum_{h=1}^{H} \eta_h^t(x_h^t) \widetilde{\ell}_h^t(x_h^t, a_h^t) \\
&\leq \frac{H}{2} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{(x_h, a_h) \in \mathcal{A}(\mathcal{X}_h)} \eta_h^t(x_h) \widetilde{p}_{1:h}^t(x_h, a_h) \\
&= \frac{\eta H}{2} \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} \frac{\widetilde{p}_{1:h}^t(x_h, a_h)}{1 + \widetilde{P}_{1:h}^{t-1}(x_h)} \\
&= \frac{\eta H}{2} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} \sum_{t=1}^{T} A_{x_h} \frac{\widetilde{p}_{1:h}^t(x_h)}{1 + \widetilde{P}_{1:h}^{t-1}(x_h)} \\
&\leq \frac{\eta H}{2} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} A_{x_h} \log_2\left(1 + \widetilde{P}_{1:h}^T(x_h)\right) \\
&= \frac{\eta H}{2} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} A_{x_h} \log_2\left(1 + \frac{1}{A_{x_h}} \sum_{a_h \in \mathcal{A}(x_h)} \widetilde{P}_{1:h}^T(x_h, a_h)\right) \\
&\leq \frac{\eta H}{2} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} A_{x_h} \log_2\left(1 + \frac{1}{A_{x_h}} \sum_{a_h \in \mathcal{A}(x_h)} (1 + 2P_{1:h}^T(x_h, a_h))\right) \\
&\leq \frac{\eta H}{2} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h} A_{x_h} \log_2(2 + 2T) \\
&= \frac{\eta \upsilon H}{2} A_{\mathcal{X}}
\end{aligned}$$

where we used $\widetilde{p}_{1:h}^t(x_h^t) \leq 1/(\gamma_h^t(x_h^t, a_h^t) A_{x_h^t}) \leq (1 + \widetilde{P}_{1:h}^{T-1}(x_h^t, a_h^t))/(\gamma A_{x_h^t}) \leq 1 + \widetilde{P}_{1:h}^{T-1}(x_h^t)$ for the assumption of Lemma D.3, and inequality 8 for the next to last inequality. □

We now have all the tools needed to upper bound the regret with high probability when using `Adaptive FTRL`.

Theorem 5.1. *Let $\iota' = \log(3A_{\mathcal{X}}T/\delta)$ and $\upsilon = 1 + \log_2(1 + T)$. For any $\delta \in (0, 1)$, using `Adaptive FTRL` with $\eta = 2\sqrt{\iota'T/(\upsilon A_{\mathcal{X}})}$, $\gamma = \sqrt{2\iota'HT/(\upsilon A_{\mathcal{X}})}$ and μ^0 the uniform policy yields with a probability at least $1 - \delta$,*

$$\mathfrak{R}_{\max}^T \leq 6H\sqrt{\iota'\upsilon A_{\mathcal{X}}T} .$$

As in Corollary 4.3, this leads to the computation of an ε-optimal profile as long as

$$T \geq \mathcal{O}\left(H^2(A_{\mathcal{X}} + B_{\mathcal{Y}}) \log(1/\delta + T) \log(T)/\varepsilon^2\right) .$$

Proof. We first notice that if $T < 4A_\mathcal{X}$, then, as $\mathfrak{R}_{\max}^T \leq T$, we have

$$\mathfrak{R}_{\max}^T < 2\sqrt{A_\mathcal{X}T}$$

and the bound is then immediate. Similarly, if $\gamma < 2\iota'$, then

$$\mathfrak{R}_{\max}^T \leq \sqrt{\frac{2\iota'\upsilon A_\mathcal{X}T}{H}}$$

and the bound also trivially holds. We will thus assume in the rest of the proof that $A_\mathcal{X} \leq T$ and $\gamma \geq 2\iota'$. This last inequality lets us use Lemma G.1 under event ε, as it also implies $\gamma \geq 1$.

In this case, using the previous lemmas, along with $2 \leq A_\mathcal{X} \leq T$, $2 \leq \iota'$, $2\log(A_\mathcal{X}) \leq \iota'$ and $2 \leq \upsilon$ for the last two inequalities:

$$
\begin{aligned}
\mathfrak{R}_{\max}^T &\leq \text{BIAS I} + \text{BIAS II} + \text{REG} + \text{VAR} \\
&\leq H\sqrt{2T\iota'} + X + \gamma\upsilon A_\mathcal{X} + \frac{2\iota'}{\gamma}HT + \frac{2\log(A_\mathcal{X})}{\eta}H(X+T) + \frac{\eta\upsilon}{2}HA_\mathcal{X} \\
&\leq (\frac{1}{\sqrt{2}} + \frac{1}{4})H\sqrt{\iota'\upsilon A_\mathcal{X}T} + \gamma\upsilon A_\mathcal{X} + \frac{2\iota'}{\gamma}HT + \frac{4\log(A_\mathcal{X})}{\eta}HT + \frac{\eta\upsilon}{2}HA_\mathcal{X} \\
&\leq H\sqrt{\iota'\upsilon A_\mathcal{X}T} + \gamma\upsilon A_\mathcal{X} + \frac{2\iota'}{\gamma}HT + \frac{2\iota'}{\eta}HT + \frac{\eta\upsilon}{2}HA_\mathcal{X}.
\end{aligned}
$$

The previous expression is minimized with $\gamma = \sqrt{2\iota'HT/(\upsilon A_\mathcal{X})}$ and $\eta = 2\sqrt{\iota'T/(\upsilon A_\mathcal{X})}$, which yields

$$
\begin{aligned}
\mathfrak{R}^T(\mu^\star) &\leq H\sqrt{\iota'\upsilon A_\mathcal{X}T} + 2H\sqrt{2\iota'\upsilon A_\mathcal{X}T} + 2H\sqrt{\iota'\upsilon A_\mathcal{X}T} \\
&\leq 6H\sqrt{\iota'\upsilon A_\mathcal{X}T}
\end{aligned}
$$

where we used $2\sqrt{2} \leq 3$.

$\square$