BEYOND KL DIVERGENCE: GENERAL PAYOFF PERTURBATIONS IN EXTENSIVE-FORM GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Payoff perturbation is a useful principle for *last-iterate convergence* in games, but existing analyzes of behavioral-form Follow-the-Regularized-Leader (FTRL) in extensive-form games have focused mainly on a KL reward transformation. We use a strongly convex potential to construct Bregman-based payoff adjustments at each history–action pair relative to a *anchoring strategy* profile, and show that these adjustments make a global Lyapunov function decrease. Under explicit regularity assumptions, this yields last-iterate convergence of continuous-time perturbed FTRL (PFTRL) to a stationary point of the modified game when the anchoring strategy profile is held fixed. We instantiate the framework with KL and reverse-KL perturbations and derive a specialized perturbation and convergence result for two-player zero-sum games. Building on anchoring strategy updates from prior work, we also analyze an idealized outer iteration that solves each modified game exactly before updating the anchoring strategy profile; this result does not cover the finite-inner-loop discrete-time implementation. Experiments on two-player zero-sum benchmarks show that discrete-time PFTRL substantially reduces last-iterate exploitability, while perturbed CFR+ variants remain the strongest solver-level baselines in most settings.

1 INTRODUCTION

Extensive-form games (EFGs) provide a standard model of sequential strategic interaction under imperfect information. A central computational task is to learn equilibrium strategies in such games (Moravčík et al., 2017; Brown & Sandholm, 2018; 2019). Counterfactual Regret Minimization (CFR) and its variants, most notably CFR+, are among the most successful practical methods for large EFGs (Zinkevich et al., 2007; Tammelin, 2014). Their classical guarantees, however, concern the time-averaged strategy; the strategy played at the current iteration need not itself converge (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018).

This distinction has motivated the study of *last-iterate convergence*. Two prominent stabilization mechanisms are optimism, which corrects an update using a prediction of the next payoff or gradient (Daskalakis et al., 2018; Daskalakis & Panageas, 2019), and payoff perturbation, which modifies the game through a strategy-dependent penalty (Perolat et al., 2021; Abe et al., 2024). For behavioral-form FTRL in EFGs, Perolat et al. (2021) establish last-iterate convergence using a Kullback–Leibler (KL) reward transformation. For CFR-type methods, Meng et al. (2025) study quadratic regularization in sequence form and develop RTCFR+, which uses local information-set updates. To parallel our Perturbed FTRL (PFTRL) terminology, we refer to their RTCFR+ as Perturbed CFR+ (PCFR+) throughout this paper.

These results leave open whether stabilization of behavioral-form FTRL is specific to KL divergence. Extending the idea to other divergences is challenging in extensive-form games: a payoff change at one decision point also affects what happens later in the game. We seek a general payoff perturbation that accounts for these downstream effects.

We construct a history–action *correction* relative to an *anchoring strategy profile*, which serves as the point of comparison. The correction combines a gradient difference with a term shared by all actions at the same decision point. This shared term is chosen so that the correction, averaged under the current strategy, equals the local Bregman divergence without changing differences between actions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We prove that these local corrections combine across the game tree to make a global Lyapunov function decrease along continuous-time PFTRL.

With the anchoring strategy profile held fixed, we prove that continuous-time PFTRL converges to a stationary point of the modified game under the assumptions in Section 4. This point need not be a Nash equilibrium of the original game. The analysis covers KL and reverse KL (RKL) perturbations, as well as a specialized construction for two-player zero-sum games. Notice that the RKL divergence is used in the context of supervised learning, it encourages the student to concentrate on outputs or actions that the teacher itself regards as highly likely (Czarnecki et al., 2019; Gu et al., 2024). It is also conceptually significant because it connects naturally to the mutation-based perspective from normal-form games (Abe et al., 2022).

Prior work proposes updating the anchoring strategy profile to recover an equilibrium of the original game (Perolat et al., 2021; Abe et al., 2023; 2024). We show that this approach converges with our general perturbation when each modified game is solved exactly before the anchoring strategy profile is updated. The practical algorithm instead takes a finite number of discrete-time steps between updates, so this outer-loop theorem does not apply directly to it. We then introduce a separate perturbation for two-player zero-sum games: it gives the players equal and opposite payoff corrections and has its own fixed-anchor convergence proof. Extending the outer-loop result to this zero-sum construction remains open.

Our experiments evaluate discrete-time PFTRL on standard two-player zero-sum benchmarks. They are intended to test whether the proposed perturbations stabilize behavioral-form FTRL, not to claim that PFTRL outperforms specialized CFR solvers. PFTRL substantially reduces last-iterate exploitability, whereas PCFR+ remains the strongest practical baseline in most settings. We also report a fully quadratic PFTRL variant whose empirical behavior is strong but whose FTRL regularizer is not Legendre and therefore lies outside our convergence theory.

2 PRELIMINARIES

We consider finite N-player extensive-form games with imperfect information in behavioral-form. An extensive-form game is a tuple $\langle N, c, H, Z, A, P, \pi_c, u, \mathcal{X}\rangle$ where N is the (finite) set of players and c is the chance player that moves by drawing random outcomes according to the fixed probability distribution $\pi_c(\cdot|h) \in \Delta(A(h))$. H is the set of histories (action sequences) with prefix order $h \sqsubseteq h'$; $Z \subseteq H$ is the set of terminal histories. For each non-terminal history $h \in H \setminus Z$, the available actions are $A(h)$ and taking $a \in A(h)$ leads to successor history ha. The player function $P : H \setminus Z \to N \cup \{c\}$ specifies the acting player at each history. We write $u_i(h, a)$ for the immediate payoff to player i when action a is taken at history h, and assume only terminal transitions yield non-zero payoff (i.e., $u_i(h, a) = 0$ if $ha \notin Z$).

For each player $i \in N$, let X_i be the collection of information sets partitioning $\{h \in H : P(h) = i\}$; let $\mathcal{X} := \bigcup_{i\in N} X_i$. For $x \in X_i$, all histories $h \in x$ share the same action set, denoted $A(x)$, and we write $x(h)$ for the information set containing h. A policy $\pi_i(\cdot \mid x) \in \Delta(A(x))$ assigns an action distribution to each $x \in X_i$, and $\pi = (\pi_i)_{i\in N}$ denotes a policy or strategy profile.

Reach probabilities. The reach probability of a history h under π is

$$\rho^\pi(h) := \prod_{(h'a')\sqsubseteq h} \pi_{P(h')}(a' \mid x(h')), \tag{1}$$

where π_c is used when $P(h') = c$. We decompose $\rho^\pi(h) = \rho_i^\pi(h)\rho_{-i}^\pi(h)$ into player i and others (including chance) contributions. For $h \sqsubseteq h'$, define the transition probability $\rho^\pi(h, h') := \rho^\pi(h')/\rho^\pi(h)$ (and 0 otherwise). We assume *perfect recall*, under which the players do not forget any information that they once observed (Zinkevich et al., 2007) and The reach probability of an information set x is $\rho^\pi(x) := \sum_{h\in x} \rho^\pi(h)$, and $\rho_i^\pi(h)$ is constant over $h \in x$ (so we write $\rho_i^\pi(x)$).

Values. Given π, define the history value at $h \in H$ as

$$v_i^\pi(h) = \sum_{h'a' \sqsupseteq h} \rho^\pi(h, h'a')u_i(h', a'). \tag{2}$$

We denote $v_i^\pi(h_{\text{init}})$ as the history value at the initial history h_{init} and define the class of games we focus on in this paper as *monotone game*, where, for any strategy profiles $\pi, \pi' \in \Sigma$,

$$v_i^{\pi_i,\pi_{-i}}(h_{\text{init}}) - v_i^{\pi'_i,\pi_{-i}}(h_{\text{init}}) - v_i^{\pi_i,\pi'_{-i}}(h_{\text{init}}) + v_i^{\pi'_i,\pi'_{-i}}(h_{\text{init}}) \leq 0 \tag{3}$$

where $\Sigma = \prod_i \Sigma_i$ symbolizes the set of all strategy profiles.

Next, define the history-action Q-value as

$$q_i^\pi(h,a) = \sum_{h'a' \sqsupseteq ha} \rho^\pi(ha, h'a') u_i(h', a'). \tag{4}$$

Finally, let us define *counterfactual value*, or value of a policy π for player i at information set x while taking action a, as

$$v_i^\pi(x,a) = \sum_{h \in x} \rho_{-i}^\pi(h) q_i^\pi(h,a). \tag{5}$$

Nash equilibrium. A popular solution concept for extensive-form games is a *Nash equilibrium* (Nash, 1951), where no player can increase their expected payoff by deviating from their designated strategy.

A Nash equilibrium $\pi^* = (\pi_i^*, \pi_{-i}^*)$ ensures the following condition: for any $i \in N$ and $\pi_i \in \Sigma_i$,

$$v_i^{\pi_i^*,\pi_{-i}^*}(h_{\text{init}}) \geq v_i^{\pi_i,\pi_{-i}^*}(h_{\text{init}}),$$

where Σ_i indicates the set of all strategies for player i. Furthermore, we define

$$\text{exploit}(\pi) := \sum_{i \in N} \left(\max_{\widetilde{\pi}_i \in \Sigma_i} v_i^{\widetilde{\pi}_i,\pi_{-i}}(h_{\text{init}}) - v_i^{\pi_i,\pi_{-i}}(h_{\text{init}}) \right),$$

as *exploitability* of a given strategy profile π, which is a metric for measuring the closeness of π to a Nash equilibrium π^*. By definition, exploitability satisfies $\text{exploit}(\pi) \geq 0$ for any π, and it is 0 if and only if π is a Nash equilibrium.

Follow-the-Regularized Leader (FTRL) We consider the online learning setting with a finite and discrete number of iterations. At each iteration $t \geq 1$, each player $i \in N$ determines her strategy π_i^t based on the previously observed counterfactual values. By traversing the entire game-tree, each player i calculates the counterfactual values $v_i^{\pi^t}(x,a)$ for each $x \in X_i$ and $a \in A(x)$.

This paper focuses on a widely used algorithm, FTRL. In imperfect information games, it defines a sequence of strategies $(\pi^t)_{t \in \{1,2,\ldots\}}$ for all $i \in N$ and $x \in \mathcal{X}$ as follows:

$$\pi_i^{t+1}(\cdot|x) = \underset{\pi \in \Delta(A(x))}{\arg\max} \left\{ \eta \left\langle \sum_{s=1}^{t} v_i^{\pi^s}(x,\cdot), \pi \right\rangle - \psi_i(\pi) \right\}, \tag{6}$$

where η is a learning rate and $\psi_i : \Delta(A(x)) \to \mathbb{R}$ is a strongly convex regularization function.

Other notations We define the KL divergence as $\text{KL}(\pi_i, \pi'_i) = \sum_j \pi_{ij} \ln \pi_{ij}/\pi'_{ij}$ for any two strategies π_i and π'_i. For a differentiable convex function f on a simplex, the Bregman divergence is $D_f(p,q) := f(p) - f(q) - \langle \nabla f(q), p - q \rangle$. We use ψ for a Legendre FTRL regularizer and write D_ψ; we also use ϕ for a perturbation potential and write D_ϕ. Note that the main notation is summarized in Appendix A.

3 PERTURBED FTRL

This section defines PFTRL for extensive-form games. Given a current strategy profile π and an anchoring strategy profile $\sigma = (\sigma_i)_{i \in N}$, PFTRL modifies the payoff attached to each history–action pair (h,a). The correction is constructed from the discrepancy between the local strategies $\pi_i(\cdot \mid x(h))$ and $\sigma_i(\cdot \mid x(h))$ under a chosen Bregman geometry; KL, reverse KL, and squared

Euclidean potentials are discussed below. We write $d_i^{\pi,\sigma}(h,a)$ for the (possibly policy-dependent) instantaneous perturbation magnitude at (h,a), and define its cumulative effect

$$\delta_i^{\pi,\sigma}(h,a) := \sum_{h'a' \sqsupseteq ha} \rho^{\pi}(ha, h'a')\, d_i^{\pi,\sigma}(h',a'). \tag{7}$$

We then obtain the perturbed Q-value for each history-action pair by summing the non-perturbed Q-value in Eq. 4 and Eq. 7 with the perturbation strength μ:

$$q_i^{\pi,\sigma}(h,a) := q_i^{\pi}(h,a) + \mu \delta_i^{\pi,\sigma}(h,a). \tag{8}$$

Following the construction of Eq. 5, we construct the perturbed counterfactual value while choosing action a at information state x under the current strategy π and the anchoring strategy σ:

$$v_i^{\pi,\sigma}(x,a) = \sum_{h \in x} \rho_{-i}^{\pi}(h) q_i^{\pi,\sigma}(h,a). \tag{9}$$

Building on those components, we have PFTRL, which updates the last-iterate strategy $\pi_i^t(\cdot \mid x)$ for each information set x at iteration t as follows:

$$\pi_i^t(\cdot \mid x) \in \operatorname*{arg\,max}_{p \in \Delta(A(x))} \left\{ \langle y^t(x), p \rangle - \psi_i(p) \right\}, \tag{10}$$

where $y_i^t(x) \in \mathbb{R}^{A(x)}$ is the cumulative score defined as

$$y_i^t(x,a) := \int_0^t v_i^{\pi^s,\sigma}(x,a)\, ds \tag{11}$$

for each $x \in X_i$ and $a \in A(x)$. Throughout, we assume that each regularizer ψ_i is Legendre and strongly convex on $\Delta(A(x))$.

Our analysis concerns the continuous-time trajectory $(\pi_i^t)_{i \in N, t \geq 0}$. Algorithm 1, by contrast, is the discrete-time procedure used in the experiments: it takes finite steps of size η and replaces the anchoring strategy profile every T_σ iterations. The fixed-anchor result in Theorem 3 does not by itself provide a convergence guarantee for this discretization. Likewise, the later anchoring-update theorem analyzes the idealized update in which every anchored perturbed game is solved exactly; a finite T_σ is a practical approximation. However, fortunately, as highlighted in Perolat et al. (2021); Abe et al. (2023; 2024), anchoring strategy updates ensure the last-iterate strategy converges to a Nash equilibrium. Note that Liu et al. (2023) shrink it at each iteration based on the current strategy profile's proximity to an underlying equilibrium so that it admits last-iterate convergence. However, it becomes challenging to choose an appropriate learning rate for the shrinking perturbation strength.

4 Theoretical analysis for monotone games

This section provides the fixed-anchor convergence guarantee for the PFTRL dynamics. A stationary point, or an equilibrium of the perturbed game, is defined as a variational inequality where unilateral deviation is not profitable and is affected by the perturbation strength $\mu \geq 0$ and the anchoring strategy profile σ.

Definition 1 (Stationary point)**.** *A strategy profile π is a stationary point if, for any $i \in N$, $x \in X_i$, and $p \in \Delta(A(x))$,*

$$\langle v_i^{\pi,\sigma}(x,\cdot),\ p - \pi_i(\cdot \mid x) \rangle \leq 0. \tag{12}$$

For given μ and σ, we refer to the associated stationary point as $\pi^{\mu,\sigma}$. It is convenient to derive an alternative representation of the variational inequality at a stationary point for extensive-form games to prove our theorems.

Lemma 2. *The stationary point $\pi^{\mu,\sigma}$ must satisfy, for any strategy profile π,*

$$\sum_{i \in N} \left(v_i^{(\pi_i, \pi_{-i}^{\mu,\sigma}),\sigma}(h_{\mathrm{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\mathrm{init}}) \right) + \tag{13}$$
$$\mu \sum_{i \in N} \sum_{h \in H} \sum_{a \in A(h)} \left(\rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(ha) - \rho^{\pi^{\mu,\sigma}}(ha) \right) d_i^{\pi^{\mu,\sigma},\sigma}(h,a) \leq 0.$$

Algorithm 1 Perturbed FTRL with anchoring strategy updates

Require: Time horizon T, learning rate η, regularization ψ_i, mutation μ, update interval T_σ

$\pi_i^1 \leftarrow \left(\frac{1}{|A(x)|}\right)_{a_i \in A_i}$ for all $i \in N$ and $x \in X_i$
$\sigma \leftarrow \pi^1$
$\kappa_i[x] \leftarrow 0$ for all $i \in N$ and $x \in X_i$
for $t = 1, 2, \ldots, T$ **do**
  for $i \in N$ **do**
    for $x \in X_i$ **do**
      Compute $v_i^{\pi^t,\sigma}(x, \cdot)$ by traversing the game tree
      $\pi_i^{t+1}(\cdot \mid x) = \arg\max_{\pi \in \Delta(A(x))} \left\{ \eta \left\langle \sum_{s=1}^t v_i^{\pi^s,\sigma}(x, \cdot), \pi \right\rangle - \psi_i(\pi) \right\}$
      $\kappa_i[x] \leftarrow \kappa_i[x] + 1$
      if $\kappa_i[x] = T_\sigma$ **then**
        $\sigma_i(\cdot|x) \leftarrow \pi_i^{t+1}(\cdot|x)$
        $\kappa_i[x] \leftarrow 0$
      end if
    end for
  end for
end for

This lemma implies that Eq. 12 is decomposed into the expected payoffs from the initial history h_{init} and the magnitude of perturbation reaching history ha.

Let us next derive the time derivative of the (summed) Bregman divergence between $\pi^{\mu,\sigma}$ and π^t, and show that it is non-positive. In normal-form games, the magnitude of the perturbation is specified via the gradient of a divergence (e.g., KL or ℓ_2) between the current strategy and an anchoring strategy. We extend this notion of perturbation to extensive-form games. In particular, at each history–action pair (h, a), we divide the history–action correction Γ by (i) the size of the information set containing h and (ii) the opponents-and-chance reach probability $\rho_{-i}^{\pi}(h)$. For the potential ϕ, define the history–action correction

$$\begin{aligned}\Gamma^{\pi,\sigma}(h,a) &:= \left[\nabla_\pi D_\phi(\pi(h), \sigma(h))\right]_a + D_\phi(\pi(h), \sigma(h)) - \langle \nabla_\pi D_\phi(\pi(h), \sigma(h)), \pi(h) \rangle \\ &= \left[\nabla\phi(\pi(h)) - \nabla\phi(\sigma(h))\right]_a + D_\phi(\pi(h), \sigma(h)) - \langle \nabla\phi(\pi(h)) - \nabla\phi(\sigma(h)), \pi(h) \rangle ,\end{aligned}$$

where $\nabla_\pi D_\phi$ denotes the gradient of the Bregman divergence with respect to π.

Theorem 3. *Fix $\mu > 0$ and an anchoring strategy profile σ. Suppose the game is a finite, perfect-recall, monotone EFG; σ and the initial PFTRL strategy lie in the relative interior of every local simplex; chance assigns positive probability to every chance action leading to a history; and a stationary point $\pi^{\mu,\sigma}$ exists in the relative interior. Let the perturbation payoff function be*

$$d_i^{\pi,\sigma}(h,a) := -\frac{\mathbf{1}[P(h) = i]}{|x(h)|\rho_{-i}^{\pi}(h)} \Gamma^{\pi,\sigma}(h,a), \tag{14}$$

where $\mathbf{1}$ is an indicator function. Then,

$$\begin{aligned}&\frac{d}{dt} \sum_{i \in N} \sum_{x \in \mathcal{X}_i} \rho_i^{\pi^{\mu,\sigma}}(x) D_\psi(\pi^{\mu,\sigma}(x), \pi^t(x)) \\ &\quad \leq -\mu \sum_{i \in N} \sum_{x \in \mathcal{X}_i} \left(\rho_i^{\pi^{\mu,\sigma}}(x) D_\phi(\pi^{\mu,\sigma}(x), \pi^t(x)) + \rho_i^{\pi^t}(x) D_\phi(\pi^t(x), \pi^{\mu,\sigma}(x)) \right).\end{aligned} \tag{15}$$

This statement implies the reach-probability weighted divergence Thus, by Lyapunov arguments (Khalil, 2015), it converges to zero, and then π^t converges to $\pi^{\mu,\sigma}$.

Note that Equation 14 contains the reach probability's reciprocal. A literal evaluation of that reciprocal can be numerically unstable when $\rho_{-i}^{\pi}(h)$ is small, but the counterfactual backup always combines the perturbation with the same opponents-and-chance reach factor. The implementation therefore computes $\rho_{-i}^{\pi}(h) d_i^{\pi,\sigma}(h,a)$ directly and never forms $1/\rho_{-i}^{\pi}(h)$.

Corollary 4. *Assume that $d_i^{\pi,\sigma}$ is defined as in Theorem 3, and let $D_\phi(p,q)$ be the KL divergence. Then*

$$d_i^{\pi,\sigma}(h,a) = -\frac{\mathbf{1}[P(h)=i]}{|x(h)|}\frac{1}{\rho_{-i}^{\pi}(h)}\ln\frac{\sigma_i(a|x(h))}{\pi_i(a|x(h))}. \tag{16}$$

In addition, assume that the regularizer is entropy $\psi_i(\pi(\cdot|x)) = \sum_{a\in A(x)}\pi_i(a|x)\ln\pi_i(a|x)$. Then, the P-FTRL trajectory converges to a stationary point exponentially fast, i.e.,

$$\mathrm{KL}(\pi^\mu,\pi^t) \le \mathrm{KL}(\pi^\mu,\pi^0)\exp\left(-\mu\xi t\right).$$

Note that this instance is slightly different from *Reward-Transformed FTRL* (Perolat et al., 2021), as we normalize the perturbation magnitude by the size of the information set.

Corollary 5. *Assume that $d_i^{\pi,\sigma}$ is defined as in Theorem 3, and let $D_\phi(p,q)$ be the reverse KL divergence. Then*

$$d_i^{\pi,\sigma}(h,a) = -\frac{\mathbf{1}[P(h)=i]}{|x(h)|}\frac{1}{\rho_{-i}^{\pi}(h)}\cdot\left(\mathrm{KL}(\sigma_i,\pi_i) + \frac{\sigma_i(a|x(h)) - \pi_i(a|x(h))}{\pi_i(a|x(h))}\right). \tag{17}$$

In addition, assume that the regularizer is entropy. Then, the PFTRL trajectory converges to a stationary point exponentially fast, i.e.,

$$\mathrm{KL}(\pi^\mu,\pi^t) \le \mathrm{KL}(\pi^\mu,\pi^0)\exp\left(-\mu\xi t\right).$$

Note that this instance is not a mere extension of previous variants of PFTRL with RKL Abe et al. (2022; 2023; 2024) for normal-form games. Beyond normalizing the perturbation magnitude by the size of the information set, the main term is also not simply the difference between the current and the anchoring strategies.

Remark 6. *This remark defines the PFTRL-L2 variant evaluated in our experiments; it is not a convergence result. Assume that $d_i^{\pi,\sigma}$ is defined as in Theorem 3, and let $D_\phi(p,q)$ be the squared-ℓ_2 divergence, i.e., $\phi(p) = \frac{1}{2}\|p\|_2^2$. Then*

$$d_i^{\pi,\sigma}(h,a) = \frac{\mathbf{1}[P(h)=i]}{|x(h)|}\frac{1}{\rho_{-i}^{\pi}(h)}\left(\sigma(a|h) - \pi(a|h) + \frac{1}{2}\left(\|\pi(h)\|_2^2 - \|\sigma(h)\|_2^2\right)\right).$$

In addition, assume that the regularizer is squared-ℓ_2 divergence, i.e., $\psi_i(\pi(\cdot|x)) = \frac{1}{2}\sum_{a\in A(x)}\pi_i(a|x)^2$.

Although this regularizer is strongly convex, it is not Legendre on the simplex. Its FTRL choice map is a Euclidean projection that may reach the boundary, so the Legendre-duality and Fenchel-coupling derivative used in our proof do not directly apply. Consequently, this instance is outside the scope of our convergence theorem.

4.1 Proof sketch of Theorem 3

We sketch below the proof of Theorem 3. The complete proof and the associated lemmas are placed in Appendix F. We first define the weighted sum of the divergence to π^t: For a given strategy profile π and for an iteration t,

$$J(t;\pi) = \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^\pi(x)D_\psi(\pi(x),\pi^t(x)).$$

Note that, for the fixed anchoring strategy σ, the Lyapunov candidate is $J(t;\pi^{\mu,\sigma})$. Since each Bregman divergence is nonnegative, we have $J(t;\pi^{\mu,\sigma}) \ge 0$, and under the strict convexity of ψ, $J(t;\pi^{\mu,\sigma}) = 0$ holds if and only if $\pi^t = \pi^{\mu,\sigma}$ at every information set. Then, we obtain and decompose the time derivative as follows:

Lemma 7. *The derivative of J in terms of time t is given as:*

$$\begin{aligned}\frac{d}{dt}J(t;\pi) &= \sum_{i\in N}\left(v_i^{(\pi^t),\sigma}(h_{\mathrm{init}}) - v_i^{(\pi_i,\pi_{-i}^t),\sigma}(h_{\mathrm{init}})\right) \\ &+ \mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho^{\pi^t}(ha) - \rho^{\pi_i,\pi_{-i}^t}(ha)\right)d_i^{\pi^t,\sigma}(h,a).\end{aligned} \tag{18}$$

This derivative identity is derived from the first-order optimality condition of the PFTRL update in Eq. 10, the Fenchel-coupling form of the Bregman divergence, and the expected payoff gain or loss incurred when an agent deviates from the current strategy π^t to an alternative strategy π, as characterized in Lemma 13 of Appendix 14.

Next, we derive the upper bound of the derivative of the Lyapunov candidate at the stationary point $\pi^{\mu,\sigma}$.

Lemma 8. *The derivative of $J(\cdot, \pi^{\mu,\sigma})$ in terms of time t is given as:*

$$\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \leq \mu \sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho_i^{\pi^{\mu,\sigma}}(ha) - \rho_i^{\pi^t}(ha)\right)\cdot\left(\rho_{-i}^{\pi^{\mu,\sigma}}(ha)d_i^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-i}^{\pi^t}(ha)d_i^{\pi^t,\sigma}(h,a)\right).$$

Lemma 8 implies that the derivative of J is bouned by the difference between the cumulative perturbation magnitude weighted by the reach probabilities, which can be rewritten as a symmetric sum of Bregman divergences.

Finally, substituting Eq. 14 into $d_i^{\pi^{\mu,\sigma},\sigma}$ and $d_i^{\pi^t,\sigma}$ in the last term of Lemma 8. Several expansion and calculation induces Eq. 15. The detail is placed in Appendix F.3. The right-hand side of Eq. 15 becomes non-positive because of the definition of Bregman divergence. This makes the standard Lyapunov argument hold, which implies $\pi^t \to \pi^{\mu,\sigma}$ as $t \to \infty$.

4.2 Convergence to an equilibrium point

Following the anchoring-update idea of prior work (Perolat et al., 2021; Abe et al., 2023; 2024), this subsection studies the idealized exact-solve map. Let σ^k denote the anchoring strategy profile after the k-th outer update. For each k, the continuous-time inner dynamics with fixed anchor σ^k are assumed to reach their stationary point π^{μ,σ^k} before the next update, and the exact outer iteration is $\sigma^{k+1} = \pi^{\mu,\sigma^k}$. This is not the literal update in Algorithm 1: after a finite T_σ, that algorithm replaces the anchor by the current discrete-time iterate, which need not equal π^{μ,σ^k}. The theorem below therefore does not guarantee convergence of the finite-T_σ implementation.

Theorem 9 (Exact-solve anchoring updates)**.** *Let the payoff perturbation be the general construction in Eq. 14. Suppose the perturbation potential ϕ is m-strongly convex and the hypotheses of Theorem 3 hold for every fixed-anchor inner problem associated with σ^k. Assume further that the selected stationary-point map $S(\sigma) := \pi^{\mu,\sigma}$ is single-valued and continuous along the iterates and that the exact-solve sequence remains in a compact subset of the relative interior. Equivalently for the argument below, assume there is a constant $\underline{\rho} > 0$ such that $\rho_i^{\sigma^{k+1}}(x) \geq \underline{\rho}$ for every relevant i, x, k. If $\sigma^{k+1} = S(\sigma^k) = \pi^{\mu,\sigma^k}$ for every k, then $\{\sigma^k\}_{k\geq 0}$ converges to a Nash equilibrium π^* of the unperturbed game; that is,*

$$\sigma^k \to \pi^* \text{ as } k \to \infty.$$

The proof of this outer-loop result uses the general perturbation in Eq. 14. The zero-sum specialization is considered separately in Section 5.

4.2.1 Proof sketch of Theorem 9

The proof has three steps. First, the next lemma gives a Bregman–Fejér descent inequality relative to every Nash equilibrium. Second, summing that inequality and using the uniform reach lower bound proves asymptotic regularity, $\|\sigma^{k+1} - \sigma^k\| \to 0$. Compactness then gives an accumulation point $\bar{\sigma}$, and continuity of the exact-solve map makes it a fixed point. At such a fixed point the perturbation vanishes, so $\bar{\sigma}$ is a Nash equilibrium of the unperturbed game. Finally, applying the same descent inequality with $\pi^* = \bar{\sigma}$ upgrades subsequential convergence to convergence of the entire sequence. Appendix H contains the full proof.

Lemma 10. *For any Nash equilibrium π^* of the underlying game, we have for any $k \geq 0$:*

$$\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)D_\phi(\pi^*(x),\sigma^{k+1}(x)) - \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)D_\phi(\pi^*(x),\sigma^{k}(x)) \tag{19}$$
$$\leq -\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\sigma^{k+1}}(x)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x)).$$

For a fixed Nash equilibrium π^*, the left-hand quantity in Lemma 10 is nonincreasing and bounded below by zero. Summing Eq. 19 over k therefore gives

$$\sum_{k=0}^{\infty}\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\sigma^{k+1}}(x)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x)) < \infty.$$

Strong convexity of ϕ and the uniform lower bound $\rho_i^{\sigma^{k+1}}(x) \geq \underline{\rho}$ imply

$$\frac{m\underline{\rho}}{2}\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\|\sigma^{k+1}(x)-\sigma^k(x)\|_2^2 \leq \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\sigma^{k+1}}(x)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x)).$$

Hence $\|\sigma^{k+1}-\sigma^k\| \to 0$. By compactness, some subsequence σ^{k_j} converges to an accumulation point $\bar\sigma$; asymptotic regularity also gives $\sigma^{k_j+1} \to \bar\sigma$. Since $\sigma^{k_j+1} = S(\sigma^{k_j})$ and S is continuous, $S(\bar\sigma) = \bar\sigma$. Substituting this fixed point into Lemma 2 gives

$$\sum_{i\in N}\left(v_i^{(\pi_i,\pi_{-i}^{\mu,\bar\sigma}),\bar\sigma}(h_{\text{init}}) - v_i^{(\pi^{\mu,\bar\sigma}),\bar\sigma}(h_{\text{init}})\right)$$
$$\leq -\mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho^{\pi_i,\pi_{-i}^{\mu,\bar\sigma}}(ha) - \rho^{\pi^{\mu,\bar\sigma}}(ha)\right)d_i^{\pi^{\mu,\bar\sigma},\bar\sigma}(h,a) = 0$$

because $d_i^{\bar\sigma,\bar\sigma}(h,a) = 0$ for every h, a. Thus $\bar\sigma$ is a Nash equilibrium. Lemma 10 with $\pi^* = \bar\sigma$ shows that the nonnegative sequence $\sum_{i,x}\rho_i^{\bar\sigma}(x)D_\phi(\bar\sigma(x),\sigma^k(x))$ is nonincreasing. It converges to zero along $\{k_j\}$ and therefore along the full sequence. Strong convexity and the positive equilibrium reach weights then yield $\sigma^k \to \bar\sigma$.

5 Two-player zero-sum specialization

For behavioral-form FTRL, we ask whether a Bregman payoff correction can exploit the zero-sum structure while allowing a general strongly convex potential. The general perturbation in Eq. 14 applies to zero-sum games, but changes only the acting player's payoff. Here we assign equal and opposite corrections to the two players at every history–action pair. This gives a distinct zero-sum-preserving transformation and a fixed-anchor Lyapunov bound with mixed-player reach weights. Its payoff-level formula has no reciprocal reach factor. Specifically, suppose $N = \{1,2\}$ and $v_1^\pi(h_{\text{init}}) = -v_2^\pi(h_{\text{init}})$ for any strategy profile π.

Theorem 11 (Two-player zero-sum games)**.** *Fix $\mu > 0$ and an anchoring strategy profile σ. Assume the game is finite, perfect-recall, and two-player zero-sum; σ and the initial PFTRL strategy lie in the relative interior of every local simplex; chance assigns positive probability to every chance action leading to a history used in the analysis; and a stationary point $\pi^{\mu,\sigma}$ exists in the relative interior. Let the perturbation payoff function be*

$$d_i^{\pi,\sigma}(h,a) := (-\mathbf{1}[P(h)=i] + \mathbf{1}[P(h)=3-i])\,\Gamma^{\pi,\sigma}(h,a). \tag{20}$$

Let $\pi^{\mu,\sigma}$ be a stationary point in Definition 1 and define $H_i := \{h \in H \setminus Z : P(h) = i\}$. Then PFTRL converges to the stationary point with the following Lyapunov decrease:

$$\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \leq -\mu\sum_{i\in N}\sum_{h\in H_i}\left(\rho^{\pi_i^{\mu,\sigma},\pi_{-i}^t}(h)\,D_\phi(\pi^{\mu,\sigma}(h),\pi^t(h)) + \rho^{\pi_i^t,\pi_{-i}^{\mu,\sigma}}(h)\,D_\phi(\pi^t(h),\pi^{\mu,\sigma}(h))\right). \tag{21}$$

The exact-solve anchoring theorem, Theorem 9, concerns Eq. 14, not the zero-sum correction above. Its Bregman–Fejér descent step uses self-reach weights that do not follow from the mixed-player reach terms in Eq. 21. Although the zero-sum correction vanishes when the strategy equals the anchor, this observation alone does not establish convergence of repeated anchor updates; that question requires a separate outer-loop argument.

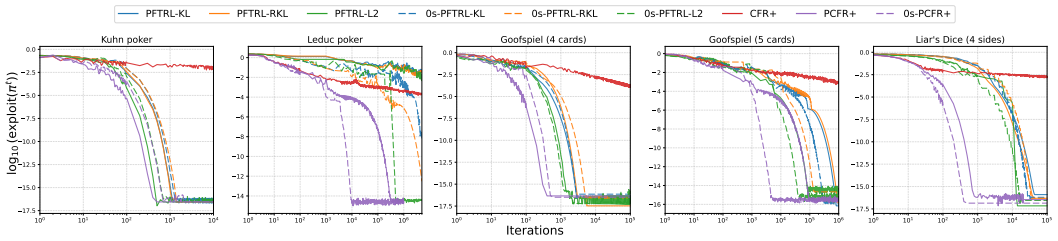


Figure 1: Exploitability of last-iterate π^t across benchmark games.

6 Experiments

In this section, we investigate how the variants of the PFTRL perform on the common benchmark games. We evaluate on Kuhn poker (Kuhn, 1951), Leduc poker (Southey et al., 2005), Goofspiel (Lanctot, 2013), and Liar's Dice (Ferguson & Ferguson, 1991). We used the OpenSpiel (Lanctot et al., 2019) framework and conducted all experiments on a machine running Windows 11 Pro 64-bit with Ubuntu on Windows Subsystem for Linux (WSL), equipped with an Intel Core i7-14700 CPU and 128 GB RAM. We compare six PFTRL instantiations—three for monotone games and three for zero-sum games—against three CFR-based baselines: unperturbed CFR+ (Tammelin, 2014), PCFR+ for monotone games, and 0s-PCFR+ for zero-sum games. The PFTRL variants PFTRL-KL, PFTRL-RKL, and PFTRL-L2 are defined in Corollaries 4, 5, and Remark 6, respectively. Their zero-sum counterparts, denoted by 0s-PFTRL-KL, 0s-PFTRL-RKL, and 0s-PFTRL-L2, are derived from Theorem 11. Figure 1 reports the last-iterate exploitability across the benchmark games. For each algorithm, we report the best-performing combination among the hyperparameters. Full hyperparameter tuning results are provided in Appendix I.

Overall, unperturbed CFR+ decreases exploitability slowly or plateaus at relatively high values, whereas PCFR+ and 0s-PCFR+ attain the lowest exploitability on most benchmarks. We construct 0s-PCFR+ by combining PCFR+ with the zero-sum perturbation in Eq. 20. Their strong performance is consistent with their specialized use of counterfactual regret decomposition and regret-matching-plus updates in extensive-form games.

Measured by iterations, the PFTRL variants generally reduce exploitability less rapidly than the PCFR+ baselines. Nevertheless, they substantially reduce last-iterate exploitability and, in several games, approach the values achieved by PCFR+ and 0s-PCFR+. Our empirical claim is therefore about the stabilizing effect of payoff perturbation on behavioral-form FTRL, not about PFTRL replacing or outperforming specialized CFR solvers. Although the finite anchor-update interval T_σ introduces inner-solve error relative to the exact-solve theory, the low exploitability observed in Figure 1 empirically supports that the practical iterates approach an equilibrium closely. The relative performance of KL and RKL perturbations is game-dependent, with neither variant uniformly dominating the other. The fully quadratic PFTRL-L2 variants are often empirically stable, despite lying outside our theoretical guarantees.

7 Conclusion

This paper introduced a divergence-based payoff perturbation framework for behavioral-form FTRL in extensive-form monotone games. Building on the KL-based reward transformation of Perolat et al. (2021), we showed that last-iterate stabilization is not tied to a single KL construction, but can be obtained through a broader class of divergence-induced perturbations. In particular, the framework yields both KL and RKL instances, and our convergence analysis establishes last-iterate convergence of the corresponding PFTRL dynamics. The experiments support the narrower empirical claim that payoff perturbation stabilizes discrete-time behavioral-form FTRL on the tested two-player zero-sum games. PCFR+ remains the stronger solver-level baseline in most settings. PFTRL-L2 also performs well empirically, but its quadratic FTRL regularizer places it outside the present convergence theory. An important question for future work is a discrete-time analysis that controls both discretization error and the finite inner-solve error created by periodic anchor updates.

REFERENCES

Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *UAI*, pp. 1–10, 2022.

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. Last-iterate convergence with full- and noisy-information feedback in two-player zero-sum games. In *AISTATS*, pp. 7999–8028, 2023.

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *ICML*, volume 235, pp. 31–80, 2024.

James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *EC*, pp. 321–338, 2018.

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. In *ICLR*, 2023.

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

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *AAAI*, pp. 1829–1836, 2019.

Wojciech M. Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant M. Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *AISTATS*, pp. 1331–1340. PMLR, 2019.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *The proceedings of the 10th Innovations in Theoretical Computer Science*, volume 124, pp. 27:1–27:18, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *ICLR*, 2018.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Gabriele Farina, Christian Kroer, Noam Brown, and Tuomas Sandholm. Stable-predictive optimistic counterfactual regret minimization. In *ICML*, volume 97, pp. 1853–1862, 2019.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *AAAI*, pp. 5363–5371, 2021.

Christopher P. Ferguson and Thomas S. Ferguson. *Models for the Game of Liar's Dice*, pp. 15–28. Springer Netherlands, 1991.

Andrew Gilpin and Tuomas Sandholm. Lossless abstraction of imperfect information games. *J. ACM*, 54(5):25–es, 2007.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *ICLR*, 2024.

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

Hassan K. Khalil. *Nonlinear Control, Global Edition*. Pearson Education, 2015.

Jayash Koshal, Angelia Nedić, and Uday V Shanbhag. Single timescale regularized stochastic approximation schemes for monotone nash games under uncertainty. In *CDC*, pp. 231–236. IEEE, 2010.

Harold W. Kuhn. *A simplified two-person poker*, pp. 97–104. Princeton University Press, 1951.

Marc Lanctot. *Monte Carlo sampling and regret minimization for equilibrium computation and decision-making in large extensive form games*. PhD thesis, University of Alberta (Canada), 2013.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arxiv preprint arXiv:1908.09453*, 2019.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. In *NeurIPS*, pp. 14293–14305, 2021.

Mingyang Liu, Asuman Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *ICLR*, 2023.

Linjian Meng, Tianpei Yang, Youzhi Zhang, Zhenxing Ge, Shangdong Yang, Tianyu Ding, Wenbin Li, Bo An, and Yang Gao. Efficient last-iterate convergence algorithms in solving games. In *NeurIPS*, 2025.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA*, pp. 2703–2717, 2018.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*, 2019.

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

John Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *ICML*, volume 139, pp. 8525–8535, 2021.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *ICLR*, 2023.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: opponent modelling in poker. In *UAI*, pp. 550–558, 2005.

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

Tatiana Tatarenko and Maryam Kamgarpour. Learning Nash equilibria in monotone games. In *CDC*, pp. 3104–3109. IEEE, 2019.

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

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR*, 2021.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. In *NeurIPS*, 2024.

Naifeng Zhang, Stephen McAleer, and Tuomas Sandholm. Faster game solving via hyperparameter schedules. *arXiv preprint arXiv:2404.09097*, 2024.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *NeurIPS*, pp. 1729–1736, 2007.

A NOTATION

Table 1: Main notation. Player and information-set subscripts are omitted when unambiguous.

Symbol	Meaning
$\boldsymbol{\pi}, \boldsymbol{\pi}^t$	behavioral strategy profile and its value at time or iteration t
$\boldsymbol{\sigma}$	anchoring strategy profile that centers the payoff perturbation
$\boldsymbol{\pi}^{\mu,\sigma}$	stationary point of the game perturbed with strength μ around σ
ϕ, D_ϕ	perturbation potential and its Bregman divergence
ψ_i, D_ψ	player i's FTRL regularizer and its Bregman divergence
$\boldsymbol{\Gamma}^{\pi,\sigma}(h,a)$	local Bregman history–action correction
$\boldsymbol{d}_i^{\pi,\sigma}(h,a)$	instantaneous payoff perturbation for player i at (h,a)
$\boldsymbol{\rho}_i^{\pi}(h)$	contribution of player i to the probability of reaching h
$\boldsymbol{\rho}_{-i}^{\pi}(h)$	opponents-and-chance contribution to reaching h; not an inverse probability
μ, η, T_σ	perturbation strength, learning rate, and anchor-update interval

B RELATED LITERATURE

Traditionally, finding equilibria in extensive-form games has involved transforming the problem into a linear programming problem, but these methods do not scale to large-scale games (Gilpin & Sandholm, 2007). CFR (Zinkevich et al., 2007) and its modern variants (Tammelin, 2014; Brown & Sandholm, 2019; Farina et al., 2019; 2021; Zhang et al., 2024) have become foundational for solving extensive-form games. They exploit the counterfactual decomposition of extensive-form games and have been extensively optimized for this domain. Thus, it is expected that CFR variants remain faster on standard benchmarks, whereas this paper studies a more general FTRL-type dynamic. Thus, the comparison should be interpreted as a trade-off between EFG-specific empirical efficiency and a general last-iterate convergence principle for FTRL. Also, Meng et al. (2025) perturb the gradient of payoffs under sequence-form parameterization, while we directly perturb them by adding perturbation magnitude.

Perturbing the payoffs with strongly convex penalties (Facchinei & Pang, 2003) has long been recognized as an effective technique for achieving last iterate convergence (Koshal et al., 2010; Tatarenko & Kamgarpour, 2019) and connects to recent advances like *Reward-Transformed FTRL* (Perolat et al., 2021), the basis of *DeepNash* (Perolat et al., 2022). This line of work has also shown strong performance in practical settings, including learning in large-scale games (Bakhtin et al., 2023) and fine-tuning large language models via preference optimization (Ye et al., 2024), often in place of optimistic algorithms.

Optimistic algorithms have shown promise in achieving last-iterate convergence, in both normal-form (Daskalakis et al., 2018; Daskalakis & Panageas, 2019; Mertikopoulos et al., 2019; Wei et al., 2021) and extensive-form (Farina et al., 2021; Lee et al., 2021; Liu et al., 2023; Sokota et al., 2023) games. The latter is connected to the *sequence-form*, not behavior-form, representation of extensive-form games (von Stengel, 1996). This viewpoint motivates geometry-aware regularizer such as dilated entropy (Hoda et al., 2010) and leads to optimistic algorithms (Farina et al., 2021; Lee et al., 2021). Furthermore, it is shown that regularization and perturbation can strengthen convergence guarantees of mirror-descent- and CFR-type algorithms (Liu et al., 2023). Our analysis focuses on behavioral-form and is not intended as a comparison with sequence-form. Moreover, we do not view payoff perturbation as a competing alternative to optimistic algorithms: payoff perturbation can in principle be combined with optimistic updates and is thus orthogonal to the use of optimism.

C GAIN (LOSS) FROM DEVIATION

This section represents gain or loss when an agent deviates from a strategy given a priori. To check whether a strategy profile is a Nash equilibrium or not, we need to verify how the expected payoffs change by deviation. We first represent the gain or loss in unperturbed games and then show that summing the local discrepancies between π and π' over all information sets exactly yields the value difference at the initial history.

Lemma 12. *For any π, π', we have:*

$$\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-\langle q_i^{\pi'}(h),\pi(h)\rangle\right)=\sum_{i\in N}\left(v_i^{\pi'}(h_{\text{init}})-v_i^{\pi_i,\pi'_{-i}}(h_{\text{init}})\right).$$

Proof. We have for any π, π':

$$\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-\langle q_i^{\pi'}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-\langle q_i^{\pi_i,\pi'_{-i}}(h),\pi(h)\rangle-\langle q_i^{\pi'}(h)-q_i^{\pi_i,\pi'_{-i}}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)-\sum_{a\in A(h)}\pi(a|h)\left(q_i^{\pi'}(h,a)-q_i^{\pi_i,\pi'_{-i}}(h,a)\right)\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)-\sum_{a\in A(h)}\pi(a|h)\left(v_i^{\pi'}(ha)-v_i^{\pi_i,\pi'_{-i}}(ha)\right)\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)\right)-\sum_{i\in N}\sum_{h\in H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(v_i^{\pi'}(ha)-v_i^{\pi_i,\pi'_{-i}}(ha)\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
0&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-\langle q_i^{\pi'}(h),\pi'(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-\langle q_i^{\pi_i,\pi'_{-i}}(h),\pi'(h)\rangle-\langle q_i^{\pi'}(h)-q_i^{\pi_i,\pi'_{-i}}(h),\pi'(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)-\sum_{a\in A(h)}\pi'(a|h)\left(q_i^{\pi'}(h,a)-q_i^{\pi_i,\pi'_{-i}}(h,a)\right)\right)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)-\sum_{a\in A(h)}\pi'(a|h)\left(v_i^{\pi'}(ha)-v_i^{\pi_i,\pi'_{-i}}(ha)\right)\right)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)\right)-\sum_{i\in N}\sum_{h\in H\setminus H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(v_i^{\pi'}(ha)-v_i^{\pi_i,\pi'_{-i}}(ha)\right).
\end{aligned}$$

Combining these:

$$\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-\langle q_i^{\pi'}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)\right)-\sum_{i\in N}\sum_{h\in H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(v_i^{\pi'}(ha)-v_i^{\pi_i,\pi'_{-i}}(ha)\right)\\
&+\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)\right)-\sum_{i\in N}\sum_{h\in H\setminus H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(v_i^{\pi'}(ha)-v_i^{\pi_i,\pi'_{-i}}(ha)\right)\\
&=\sum_{i\in N}\sum_{h\in H}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi'}(h)-v_i^{\pi_i,\pi'_{-i}}(h)\right)-\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(v_i^{\pi'}(ha)-v_i^{\pi_i,\pi'_{-i}}(ha)\right)\\
&=\sum_{i\in N}\left(v_i^{\pi'}(h_{\text{init}})-v_i^{\pi_i,\pi'_{-i}}(h_{\text{init}})\right).
\end{aligned}$$

□

Next, we show how the same gain (loss) relation extends to the perturbed games. Compared with Lemma 12, the difference in the original game is preserved, and an additional correction term appears that explicitly captures the effect of the perturbation through $d_i^{\pi',\sigma}$.

Lemma 13. *Using the notation in the main text, let*

$$v_i^{\pi,\sigma}(h) = \sum_{a \in A(h)} \pi(a|h) q_i^{\pi,\sigma}(h,a).$$

Then, for any π, π', we have:

$$\begin{aligned}
&\sum_{i \in N} \sum_{x \in \mathcal{X}_i} \sum_{h \in x} \rho^{\pi_i, \pi'_{-i}}(h) \left(v_i^{\pi',\sigma}(h) - \langle q_i^{\pi',\sigma}(h), \pi(h) \rangle \right) \\
&= \sum_{i \in N} \left(v_i^{\pi'}(h_{\text{init}}) - v_i^{\pi_i, \pi'_{-i}}(h_{\text{init}}) \right) + \mu \sum_{i \in N} \sum_{h \in H} \sum_{a \in A(h)} \left(\rho^{\pi'}(ha) - \rho^{\pi_i, \pi'_{-i}}(ha) \right) d_i^{\pi',\sigma}(h,a).
\end{aligned}$$

Proof. Let us define

$$v_i^{d,\pi}(h) := \sum_{a \in A(h)} \pi(a|h) \delta_i^{\pi,\sigma}(h,a).$$

Using these functions, we can write the perturbed value functions as:

$$\begin{aligned}
q_i^{\pi,\sigma}(h,a) &= q_i^{\pi}(h,a) + \mu \delta_i^{\pi,\sigma}(h,a), \\
v_i^{\pi,\sigma}(h) &= v_i^{\pi}(h) + \mu v_i^{d,\pi}(h).
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
&\sum_{i \in N} \sum_{x \in \mathcal{X}_i} \sum_{h \in x} \rho^{\pi_i, \pi'_{-i}}(h) \left(v_i^{\pi',\sigma}(h) - \langle q_i^{\pi',\sigma}(h), \pi(h) \rangle \right) \\
&= \sum_{i \in N} \sum_{h \in H_i} \rho^{\pi_i, \pi'_{-i}}(h) \left(v_i^{\pi',\sigma}(h) - \langle q_i^{\pi',\sigma}(h), \pi(h) \rangle \right) \\
&= \sum_{i \in N} \sum_{h \in H_i} \rho^{\pi_i, \pi'_{-i}}(h) \left(v_i^{\pi'}(h) - \langle q_i^{\pi'}(h), \pi(h) \rangle \right) + \mu \sum_{i \in N} \sum_{h \in H_i} \rho^{\pi_i, \pi'_{-i}}(h) \left(v_i^{d,\pi'}(h) - \langle \delta_i^{\pi',\sigma}(h), \pi(h) \rangle \right).
\end{aligned} \tag{22}$$

From Lemma 12, the first term in equation 22 is equivalent to:

$$\sum_{i \in N} \sum_{x \in \mathcal{X}_i} \sum_{h \in x} \rho^{\pi_i, \pi'_{-i}}(h) \left(v_i^{\pi'}(h) - \langle q_i^{\pi'}(h), \pi(h) \rangle \right) = \sum_{i \in N} \left(v_i^{\pi'}(h_{\text{init}}) - v_i^{\pi_i, \pi'_{-i}}(h_{\text{init}}) \right). \tag{23}$$

In terms of the second term in equation 22, we have:

$$
\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-\langle\delta_i^{\pi',\sigma}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-\langle\delta_i^{\pi_i,\pi'_{-i},\sigma}(h),\pi(h)\rangle-\langle\delta_i^{\pi',\sigma}(h)-\delta_i^{\pi_i,\pi'_{-i},\sigma}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)-\sum_{a\in A(h)}\pi(a|h)\left(\delta_i^{\pi',\sigma}(h,a)-\delta_i^{\pi_i,\pi'_{-i},\sigma}(h,a)\right)\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\Bigg(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)\\
&\qquad\qquad-\sum_{a\in A(h)}\pi(a|h)\left(d_i^{\pi',\sigma}(h,a)+v_i^{d,\pi'}(ha)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)-v_i^{d,\pi_i,\pi'_{-i}}(ha)\right)\Bigg)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)\right)\\
&\quad-\sum_{i\in N}\sum_{h\in H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(d_i^{\pi',\sigma}(h,a)+v_i^{d,\pi'}(ha)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)-v_i^{d,\pi_i,\pi'_{-i}}(ha)\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
0&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-\langle\delta_i^{\pi',\sigma}(h),\pi'(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-\langle\delta_i^{\pi_i,\pi'_{-i},\sigma}(h),\pi'(h)\rangle-\langle\delta_i^{\pi',\sigma}(h)-\delta_i^{\pi_i,\pi'_{-i},\sigma}(h),\pi'(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)-\sum_{a\in A(h)}\pi'(a|h)\left(\delta_i^{\pi',\sigma}(h,a)-\delta_i^{\pi_i,\pi'_{-i},\sigma}(h,a)\right)\right)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\Bigg(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)\\
&\qquad\qquad-\sum_{a\in A(h)}\pi'(a|h)\left(d_i^{\pi',\sigma}(h,a)+v_i^{d,\pi'}(ha)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)-v_i^{d,\pi_i,\pi'_{-i}}(ha)\right)\Bigg)\\
&=\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)\right)\\
&\quad-\sum_{i\in N}\sum_{h\in H\setminus H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(d_i^{\pi',\sigma}(h,a)+v_i^{d,\pi'}(ha)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)-v_i^{d,\pi_i,\pi'_{-i}}(ha)\right).
\end{aligned}
$$

Combining these:

$$
\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-\langle\delta_i^{\pi',\sigma}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\sum_{h\in H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)\right)\\
&\quad-\sum_{i\in N}\sum_{h\in H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(d_i^{\pi',\sigma}(h,a)+v_i^{d,\pi'}(ha)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)-v_i^{d,\pi_i,\pi'_{-i}}(ha)\right)\\
&\quad+\sum_{i\in N}\sum_{h\in H\setminus H_i}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)\right)\\
&\quad-\sum_{i\in N}\sum_{h\in H\setminus H_i}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(d_i^{\pi',\sigma}(h,a)+v_i^{d,\pi'}(ha)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)-v_i^{d,\pi_i,\pi'_{-i}}(ha)\right)\\
&=\sum_{i\in N}\sum_{h\in H}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{d,\pi'}(h)-v_i^{d,\pi_i,\pi'_{-i}}(h)\right)\\
&\quad-\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(d_i^{\pi',\sigma}(h,a)+v_i^{d,\pi'}(ha)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)-v_i^{d,\pi_i,\pi'_{-i}}(ha)\right)\\
&=\sum_{i\in N}\left(v_i^{d,\pi'}(h_{\text{init}})-v_i^{d,\pi_i,\pi'_{-i}}(h_{\text{init}})\right)-\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(d_i^{\pi',\sigma}(h,a)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)\right).
\end{aligned}
\tag{24}
$$

Combining equation 22, equation 23, and equation 24, we have:

$$
\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi',\sigma}(h)-\langle q_i^{\pi',\sigma}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\left(v_i^{\pi'}(h_{\text{init}})-v_i^{\pi_i,\pi'_{-i}}(h_{\text{init}})\right)\\
&\quad+\mu\sum_{i\in N}\left(v_i^{d,\pi'}(h_{\text{init}})-v_i^{d,\pi_i,\pi'_{-i}}(h_{\text{init}})\right)-\mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)\left(d_i^{\pi',\sigma}(h,a)-d_i^{\pi_i,\pi'_{-i},\sigma}(h,a)\right).
\end{aligned}
$$

Here, we have:

$$
\begin{aligned}
v_i^{d,\pi'}(h_{\text{init}})&=\sum_{h\in H}\sum_{a\in A(h)}\rho^{\pi'}(ha)d_i^{\pi',\sigma}(h,a),\\
v_i^{d,\pi_i,\pi'_{-i}}(h_{\text{init}})&=\sum_{h\in H}\sum_{a\in A(h)}\rho^{\pi_i,\pi'_{-i}}(ha)d_i^{\pi_i,\pi'_{-i},\sigma}(h,a).
\end{aligned}
$$

Therefore, we finally get:

$$
\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi'_{-i}}(h)\left(v_i^{\pi',\sigma}(h)-\langle q_i^{\pi',\sigma}(h),\pi(h)\rangle\right)\\
&=\sum_{i\in N}\left(v_i^{\pi'}(h_{\text{init}})-v_i^{\pi_i,\pi'_{-i}}(h_{\text{init}})\right)+\mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho^{\pi'}(ha)-\rho^{\pi_i,\pi'_{-i}}(ha)\right)d_i^{\pi',\sigma}(h,a).
\end{aligned}
$$

□

D Proof of Lemma 2

Proof. From Eq. 12, we have for any π:

$$
\rho_i^{\pi}(x)\left\langle\sum_{h\in x}\rho_{-i}^{\pi^{\mu,\sigma}}(h)q_i^{\pi^{\mu,\sigma},\sigma}(h),\pi(x)-\pi^{\mu,\sigma}(x)\right\rangle=\sum_{h\in x}\rho^{\pi_i,\pi_{-i}^{\mu,\sigma}}(h)\left(\langle q_i^{\pi^{\mu,\sigma},\sigma}(h),\pi(h)\rangle-v_i^{\pi^{\mu,\sigma},\sigma}(h)\right)\le 0.
$$

Summing up this inequality among $i \in N$ and $x \in \mathcal{X}_i$, we have:

$$\sum_{i \in N} \sum_{x \in \mathcal{X}_i} \sum_{h \in x} \rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(h) \left(\langle q_i^{\pi^{\mu,\sigma},\sigma}(h), \pi(h) \rangle - v_i^{\pi^{\mu,\sigma},\sigma}(h) \right) \leq 0.$$

Here, from Lemma 13, the left hand side is equivalent to:

$$\sum_{i \in N} \sum_{x \in \mathcal{X}_i} \sum_{h \in x} \rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(h) \left(\langle q_i^{\pi^{\mu,\sigma},\sigma}(h), \pi(h) \rangle - v_i^{\pi^{\mu,\sigma},\sigma}(h) \right)$$
$$= \sum_{i \in N} \left(v_i^{\pi_i, \pi_{-i}^{\mu,\sigma}}(h_{\text{init}}) - v_i^{\pi^{\mu,\sigma}}(h_{\text{init}}) \right) + \mu \sum_{i \in N} \sum_{h \in H} \sum_{a \in A(h)} \left(\rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(ha) - \rho^{\pi^{\mu,\sigma}}(ha) \right) d_i^{\pi^{\mu,\sigma},\sigma}(h, a).$$

Combining these inequalities, we finally obtain:

$$\sum_{i \in N} \left(v_i^{\pi_i, \pi_{-i}^{\mu,\sigma}}(h_{\text{init}}) - v_i^{\pi^{\mu,\sigma}}(h_{\text{init}}) \right) + \mu \sum_{i \in N} \sum_{h \in H} \sum_{a \in A(h)} \left(\rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(ha) - \rho^{\pi^{\mu,\sigma}}(ha) \right) d_i^{\pi^{\mu,\sigma},\sigma}(h, a) \leq 0.$$

□

E Proof of Lemma 14

Lemma 14. *The strategy $\pi^{\mu,\sigma}$ must satisfy: for any π:*

$$\sum_{i \in N} \left(v_i^{(\pi_i, \pi_{-i}^{\mu,\sigma}),\sigma}(h_{\text{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\text{init}}) \right) - \mu \sum_{i \in N} \sum_{x \in \mathcal{X}_i} \sum_{a \in A(x)} \left(\rho_i^{\pi}(x)\pi(a|x) - \rho_i^{\pi^{\mu,\sigma}}(x)\pi^{\mu,\sigma}(a|x) \right) \cdot$$
$$\left(\frac{\partial \phi(\pi^{\mu,\sigma}(x))}{\partial \pi(a|x)} - \frac{\partial \phi(\sigma(x))}{\partial \pi(a|x)} + D_\phi(\pi^{\mu,\sigma}(x), \sigma(x)) - \langle \nabla \phi(\pi^{\mu,\sigma}(x)) - \nabla \phi(\sigma(x)), \pi^{\mu,\sigma}(x) \rangle \right) \leq 0.$$

Proof. From the definition of the function $d_i^{\pi,\sigma}(h, a)$ and Lemma 2, we have for any π:

$$\sum_{i \in N} \left(v_i^{(\pi_i, \pi_{-i}^{\mu,\sigma}),\sigma}(h_{\text{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\text{init}}) \right) + \mu \sum_{i \in N} \sum_{h \in H} \sum_{a \in A(h)} \left(\rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(ha) - \rho^{\pi^{\mu,\sigma}}(ha) \right) d_i^{\pi^{\mu,\sigma},\sigma}(h, a)$$
$$= \sum_{i \in N} \left(v_i^{(\pi_i, \pi_{-i}^{\mu,\sigma}),\sigma}(h_{\text{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\text{init}}) \right) + \mu \sum_{i \in N} \sum_{h \in H} \sum_{a \in A(h)} \left(\rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(ha) - \rho^{\pi^{\mu,\sigma}}(ha) \right) \cdot$$
$$\left(-\frac{\mathbf{1}[P(h) = i]}{|x(h)| \rho_{-i}^{\pi^{\mu,\sigma}}(h)} \left(\frac{\partial \phi(\pi^{\mu,\sigma}(h))}{\partial \pi(a|h)} - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} + D_\phi(\pi^{\mu,\sigma}(h), \sigma(h)) - \langle \nabla \phi(\pi^{\mu,\sigma}(h)) - \nabla \phi(\sigma(h)), \pi^{\mu,\sigma}(h) \rangle \right) \right)$$
$$\leq 0.$$

Here, because of the identifier function $\mathbf{1}[P(h) = i]$, the summation over $h \in H$ can be replaced by the summation over $h \in H_i$. Thus, we transform the above equation as:

$$\sum_{i \in N} \left(v_i^{(\pi_i, \pi_{-i}^{\mu,\sigma}),\sigma}(h_{\text{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\text{init}}) \right) - \mu \sum_{i \in N} \sum_{h \in H_i} \sum_{a \in A(h)} \left(\rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(ha) - \rho^{\pi^{\mu,\sigma}}(ha) \right) \cdot$$
$$\left(\frac{1}{|x(h)| \rho_{-i}^{\pi^{\mu,\sigma}}(h)} \left(\frac{\partial \phi(\pi^{\mu,\sigma}(h))}{\partial \pi(a|h)} - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} + D_\phi(\pi^{\mu,\sigma}(h), \sigma(h)) - \langle \nabla \phi(\pi^{\mu,\sigma}(h)) - \nabla \phi(\sigma(h)), \pi^{\mu,\sigma}(h) \rangle \right) \right),$$

Here, for each $h \in H_i$, we have

$$\rho^{\pi_i, \pi_{-i}^{\mu,\sigma}}(ha) = \rho_i^{\pi}(x(h))\pi(a \mid h)\rho_{-i}^{\pi_{-i}^{\mu,\sigma}}(h), \quad \rho^{\pi^{\mu,\sigma}}(ha) = \rho_i^{\pi^{\mu,\sigma}}(x(h))\pi^{\mu,\sigma}(a \mid h)\rho_{-i}^{\pi_{-i}^{\mu,\sigma}}(h).$$

Therefore, the factor $\rho_{-i}^{\pi_{-i}^{\mu,\sigma}}(h)$ appears in the numerator of both terms and cancels with the denominator $\rho_{-i}^{\pi^{\mu,\sigma}}(h)$. Moreover, when we group the summation over $h \in H_i$ according to $x = x(h)$, each information set x appears exactly $|x|$ times, which cancels with the denominator $|x(h)| = |x|$. Hence, the above equation can be rewritten as:

$$\sum_{i \in N} \left(v_i^{(\pi_i, \pi_{-i}^{\mu,\sigma}),\sigma}(h_{\text{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\text{init}}) \right) - \mu \sum_{i \in N} \sum_{x \in \mathcal{X}_i} \sum_{a \in A(x)} \left(\rho_i^{\pi}(x)\pi(a|x) - \rho_i^{\pi^{\mu,\sigma}}(x)\pi^{\mu,\sigma}(a|x) \right) \cdot$$
$$\left(\frac{\partial \phi(\pi^{\mu,\sigma}(x))}{\partial \pi(a|x)} - \frac{\partial \phi(\sigma(x))}{\partial \pi(a|x)} + D_\phi(\pi^{\mu,\sigma}(x), \sigma(x)) - \langle \nabla \phi(\pi^{\mu,\sigma}(x)) - \nabla \phi(\sigma(x)), \pi^{\mu,\sigma}(x) \rangle \right).$$

Thus, we have:

$$\sum_{i\in N}\left(v_i^{(\pi_i,\pi_{-i}^{\mu,\sigma}),\sigma}(h_{\text{init}})-v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\text{init}})\right)-\mu\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\pi}(x)\pi(a|x)-\rho_i^{\pi^{\mu,\sigma}}(x)\pi^{\mu,\sigma}(a|x)\right)\cdot$$
$$\left(\frac{\partial\phi(\pi^{\mu,\sigma}(x))}{\partial\pi(a|x)}-\frac{\partial\phi(\sigma(x))}{\partial\pi(a|x)}+D_\phi(\pi^{\mu,\sigma}(x),\sigma(x))-\langle\nabla\phi(\pi^{\mu,\sigma}(x))-\nabla\phi(\sigma(x)),\pi^{\mu,\sigma}(x)\rangle\right)\leq 0.$$

□

F Proof of Theorem 3

F.1 Proof of Lemma 7

Proof of Lemma 7. The derivative of J in terms of time t is given as:

$$\begin{aligned}\frac{d}{dt}J(t;\pi)&=\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\frac{d}{dt}D_\psi(\pi(x),\pi^t(x))\\&=\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\frac{d}{dt}\left(\psi(\pi(x))-\psi(\pi^t(x))-\left\langle\nabla\psi(\pi^t(x)),\pi(x)-\pi^t(x)\right\rangle\right)\\&=\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\frac{d}{dt}\left(-\psi(\pi^t(x))-\left\langle\nabla\psi(\pi^t(x)),\pi(x)-\pi^t(x)\right\rangle\right).\end{aligned}$$

Under the assumption that ψ is Legendre, from the first-order optimality condition for $\pi^t(x)$, there exists $\lambda(x)\in\mathbb{R}$ such that:

$$y^t(x)-\nabla\psi(\pi^t(x))=\lambda(x)\mathbf{1}. \tag{25}$$

This is the KKT condition for maximizing $\langle y^t(x),p\rangle-\psi(p)$ over the simplex $\Delta(A(x))$. The multiplier $\lambda(x)$ appears because of the simplex constraint $\sum_{a\in A(x)}p(a)=1$. Since both $\pi(x)$ and $\pi^t(x)$ lie in the simplex, we have $\langle\mathbf{1},\pi(x)-\pi^t(x)\rangle=0$, so the multiplier term vanishes after pairing with $\pi(x)-\pi^t(x)$. Therefore,

$$\langle y^t(x),\pi(x)-\pi^t(x)\rangle=\langle\nabla\psi(\pi^t(x)),\pi(x)-\pi^t(x)\rangle. \tag{26}$$

Substituting equation 26 into the last expression for $\frac{d}{dt}J(t;\pi)$, we obtain

$$\begin{aligned}\frac{d}{dt}J(t;\pi)&=\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\frac{d}{dt}\left(-\psi(\pi^t(x))-\left\langle y^t(x),\pi(x)-\pi^t(x)\right\rangle\right)\\&=\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\frac{d}{dt}\left(\langle y^t(x),\pi^t(x)\rangle-\psi(\pi^t(x))-\left\langle y^t(x),\pi(x)\right\rangle\right).\end{aligned}$$

Now we use the Legendre transform. Since $\pi^t(x)$ is the unique maximizer of $\langle y^t(x),p\rangle-\psi(p)$ over $\Delta(A(x))$,

$$\psi^*(y^t(x))=\max_{p\in\Delta(A(x))}\{\langle y^t(x),p\rangle-\psi(p)\}=\langle y^t(x),\pi^t(x)\rangle-\psi(\pi^t(x)). \tag{27}$$

Using equation 27, the previous expression becomes

$$\frac{d}{dt}J(t;\pi)=\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\frac{d}{dt}\left(\psi^*(y^t(x))-\left\langle y^t(x),\pi(x)\right\rangle\right).$$

Thus, we have:

$$\begin{aligned}
\frac{d}{dt}J(t;\pi) &= \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\left\langle \frac{d}{dt}y^t(x), \nabla\psi^*(y^t(x)) - \pi(x)\right\rangle \\
&= \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\left\langle \frac{d}{dt}y^t(x), \pi^t(x) - \pi(x)\right\rangle \\
&= \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi}(x)\left\langle \sum_{h\in x}\rho_{-i}^{\pi^t}(h)q_i^{\pi^t,\sigma}(h), \pi^t(x) - \pi(x)\right\rangle \\
&= \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi_{-i}^t}(h)\left\langle q_i^{\pi^t,\sigma}(h), \pi^t(x) - \pi(x)\right\rangle \\
&= \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{h\in x}\rho^{\pi_i,\pi_{-i}^t}(h)\left(v_i^{\pi^t,\sigma}(h) - \left\langle q_i^{\pi^t,\sigma}(h), \pi(h)\right\rangle\right).
\end{aligned}$$

Here the first equality uses the chain rule for ψ^*, the second uses $\nabla\psi^*(y^t(x)) = \pi^t(x)$, which follows from Legendre duality, the third substitutes the definition of the score dynamics $\dot{y}^t(x)$, the fourth uses $\rho_i^{\pi}(x)\rho_{-i}^{\pi^t}(h) = \rho^{\pi_i,\pi_{-i}^t}(h)$ for $h \in x$, and the last equality uses $v_i^{\pi^t,\sigma}(h) = \langle q_i^{\pi^t,\sigma}(h), \pi^t(h)\rangle$.

Finally, from Lemma 13, we have:

$$\frac{d}{dt}J(t;\pi) = \sum_{i\in N}\left(v_i^{(\pi^t),\sigma}(h_{\mathrm{init}}) - v_i^{(\pi_i,\pi_{-i}^t),\sigma}(h_{\mathrm{init}})\right) + \mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho^{\pi^t}(ha) - \rho^{\pi_i,\pi_{-i}^t}(ha)\right)d_i^{\pi^t,\sigma}(h,a)$$

□

F.2 Proof of Lemma 8

Proof. To derive the uppper bound, we first derive the inequality of the derivative of J by using the monotonicity of the expected payoff, which induces

$$\sum_{i\in N}\left(v_i^{(\pi^t),\sigma}(h_{\mathrm{init}}) - v_i^{(\pi_i^{\mu,\sigma},\pi_{-i}^t),\sigma}(h_{\mathrm{init}})\right) \le \sum_{i\in N}\left(v_i^{(\pi_i^t,\pi_{-i}^{\mu,\sigma}),\sigma}(h_{\mathrm{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\mathrm{init}})\right).$$

We substitute this inequality into Eq. 18 and obtain

$$\begin{aligned}
\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \le{}& \sum_{i\in N}\left(v_i^{(\pi_i^t,\pi_{-i}^{\mu,\sigma}),\sigma}(h_{\mathrm{init}}) - v_i^{(\pi^{\mu,\sigma}),\sigma}(h_{\mathrm{init}})\right) \\
&+ \mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho^{\pi^t}(ha) - \rho^{\pi_i^{\mu,\sigma},\pi_{-i}^t}(ha)\right)d_i^{\pi^t,\sigma}(h,a).
\end{aligned}$$

Futhermore, substituting the property of the stationary point $\pi^{\mu,\sigma}$ in Eq. 13 into the first term of the right hand side, we obtain

$$\begin{aligned}
&\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \\
&\le \mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho^{\pi^{\mu,\sigma}}(ha) - \rho^{\pi_i^t,\pi_{-i}^{\mu,\sigma}}(ha)\right)d_i^{\pi^{\mu,\sigma},\sigma}(h,a) \\
&\quad + \mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho^{\pi^t}(ha) - \rho^{\pi_i^{\mu,\sigma},\pi_{-i}^t}(ha)\right)d_i^{\pi^t,\sigma}(h,a) \\
&= \mu\sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)}\left(\rho_i^{\pi^{\mu,\sigma}}(ha) - \rho_i^{\pi^t}(ha)\right)\cdot\left(\rho_{-i}^{\pi^{\mu,\sigma}}(ha)d_i^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-i}^{\pi^t}(ha)d_i^{\pi^t,\sigma}(h,a)\right).
\end{aligned}$$

The proof of Lemma 8 is complete. □

F.3 THE REST OF THE PROOF OF THEOREM 3

Proof. From Lemma 8, we have:

$$
\begin{aligned}
&\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \\
&\leq \mu \sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)} \left(\rho_i^{\pi^{\mu,\sigma}}(ha) - \rho_i^{\pi^t}(ha)\right)\left(\rho_{-i}^{\pi^{\mu,\sigma}}(ha) d_i^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-i}^{\pi^t}(ha) d_i^{\pi^t,\sigma}(h,a)\right) \\
&= -\mu \sum_{i\in N}\sum_{h\in H_i}\sum_{a\in A(h)} \left(\rho_i^{\pi^{\mu,\sigma}}(ha) - \rho_i^{\pi^t}(ha)\right)\frac{1}{|x(h)|} \\
&\quad \cdot \left(\left[\frac{\partial}{\partial\pi(a|h)}\phi(\pi^\mu(h)) - \frac{\partial}{\partial\pi(a|h)}\phi(\sigma(h)) + D_\phi(\pi^\mu(h),\sigma(h)) - \left\langle \nabla\phi(\pi^\mu(h)) - \nabla\phi(\sigma(h)),\, \pi^\mu(h)\right\rangle\right]\right. \\
&\quad \left. - \left[\frac{\partial}{\partial\pi(a|h)}\phi(\pi^t(h)) - \frac{\partial}{\partial\pi(a|h)}\phi(\sigma(h)) + D_\phi(\pi^t(h),\sigma(h)) - \left\langle \nabla\phi(\pi^t(h)) - \nabla\phi(\sigma(h)),\, \pi^t(h)\right\rangle\right]\right).
\end{aligned}
$$

Here we substituted the definition of $d_i^{\pi,\sigma}(h,a)$. Because $d_i^{\pi,\sigma}(h,a)$ contains the indicator $\mathbf{1}[P(h) = i]$ and the factor $1/\rho_{-i}^{\pi}(h)$, the reach-weighted expression reduces to a sum over $h \in H_i$, leaving only the player-i reach probabilities.

Next we expand the Bregman divergence inside each bracket:

$$
\begin{aligned}
\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \leq -\mu \sum_{i\in N}\sum_{h\in H_i}\sum_{a\in A(h)} &\left(\rho_i^{\pi^{\mu}}(ha) - \rho_i^{\pi^t}(ha)\right)\frac{1}{|x(h)|} \\
&\cdot \left(\left[\frac{\partial}{\partial\pi(a|h)}\phi(\pi^\mu(h)) - \frac{\partial}{\partial\pi(a|h)}\phi(\sigma(h))\right.\right. \\
&\quad + \phi(\pi^\mu(h)) - \phi(\sigma(h)) - \langle\nabla\phi(\sigma(h)), \pi^\mu(h) - \sigma(h)\rangle \\
&\quad \left. - \langle\nabla\phi(\pi^\mu(h)) - \nabla\phi(\sigma(h)), \pi^\mu(h)\rangle\right] \\
&\quad - \left[\frac{\partial}{\partial\pi(a|h)}\phi(\pi^t(h)) - \frac{\partial}{\partial\pi(a|h)}\phi(\sigma(h))\right. \\
&\quad + \phi(\pi^t(h)) - \phi(\sigma(h)) - \langle\nabla\phi(\sigma(h)), \pi^t(h) - \sigma(h)\rangle \\
&\quad \left.\left. - \langle\nabla\phi(\pi^t(h)) - \nabla\phi(\sigma(h)), \pi^t(h)\rangle\right]\right).
\end{aligned}
$$

After collecting terms, the quantities depending only on $\sigma(h)$ cancel between the two brackets:

$$
\begin{aligned}
\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \leq -\mu \sum_{i\in N}\sum_{h\in H_i}\sum_{a\in A(h)} &\left(\rho_i^{\pi^{\mu}}(ha) - \rho_i^{\pi^t}(ha)\right)\frac{1}{|x(h)|} \\
&\cdot \left(\left[\frac{\partial}{\partial\pi(a|h)}\phi(\pi^\mu(h)) + \phi(\pi^\mu(h)) - \langle\nabla\phi(\pi^\mu(h)), \pi^\mu(h)\rangle\right]\right. \\
&\quad \left. - \left[\frac{\partial}{\partial\pi(a|h)}\phi(\pi^t(h)) + \phi(\pi^t(h)) - \langle\nabla\phi(\pi^t(h)), \pi^t(h)\rangle\right]\right).
\end{aligned}
$$

The remaining bracketed term depends only on the information set $x(h)$. Therefore the normalization by $|x(h)|$ allows us to aggregate the sum over histories in the same information set:

$$
\begin{aligned}
\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \leq -\mu \sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)} &\left(\rho_i^{\pi^{\mu}}(x)\pi^\mu(a|x) - \rho_i^{\pi^t}(x)\pi^t(a|x)\right) \\
&\cdot \left(\frac{\partial}{\partial\pi(a|x)}\phi(\pi^\mu(x)) + \phi(\pi^\mu(x)) - \langle\nabla\phi(\pi^\mu(x)), \pi^\mu(x)\rangle\right. \\
&\quad \left. - \frac{\partial}{\partial\pi(a|x)}\phi(\pi^t(x)) - \phi(\pi^t(x)) + \langle\nabla\phi(\pi^t(x)), \pi^t(x)\rangle\right).
\end{aligned}
$$

Finally, we rearrange the sum action-wise and recognize the two resulting terms as Bregman divergences:

$$\begin{aligned}\frac{d}{dt}J(t;\pi^{\mu,\sigma}) &\leq -\mu \sum_{i\in N}\sum_{x\in\mathcal{X}_i} \rho_i^{\pi^\mu}(x)\left(\phi(\pi^\mu(x)) - \phi(\pi^t(x)) - \langle \nabla\phi(\pi^t(x)), \pi^\mu(x) - \pi^t(x)\rangle\right) \\ &\quad + \mu \sum_{i\in N}\sum_{x\in\mathcal{X}_i} \rho_i^{\pi^t}(x)\left(\phi(\pi^\mu(x)) - \phi(\pi^t(x)) + \langle \nabla\phi(\pi^\mu(x)), \pi^t(x) - \pi^\mu(x)\rangle\right) \\ &= -\mu \sum_{i\in N}\sum_{x\in\mathcal{X}_i} \left(\rho_i^{\pi^\mu}(x) D_\phi(\pi^\mu(x), \pi^t(x)) + \rho_i^{\pi^t}(x) D_\phi(\pi^t(x), \pi^\mu(x))\right).\end{aligned}$$

□

G PROOF OF THEOREM 11

Proof. Let us define

$$f^{\pi,\sigma} := \left(\frac{\partial}{\partial\pi(a|h)}\phi(\pi(h)) - \frac{\partial}{\partial\pi(a|h)}\phi(\sigma(h)) + D_\phi(\pi(h),\sigma(h)) - \langle\nabla\phi(\pi(h)) - \nabla\phi(\sigma(h)), \pi(h)\rangle\right)$$

From Lemma 8, we have the following bound. Since we restrict attention to the two-player case $N = \{1, 2\}$, the outer summation over $i \in N$ can be expanded into the two terms corresponding to $i = 1$ and $i = 2$:

$$\begin{aligned}&\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \\ &\leq \mu \sum_{i\in N}\sum_{h\in H}\sum_{a\in A(h)} \left(\rho_i^{\pi^{\mu,\sigma}}(ha) - \rho_i^{\pi^t}(ha)\right)\left(\rho_{-i}^{\pi^{\mu,\sigma}}(ha) d_i^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-i}^{\pi^t}(ha) d_i^{\pi^t,\sigma}(h,a)\right) \\ &= \mu \sum_{h\in H}\sum_{a\in A(h)} \left(\rho_1^{\pi^{\mu,\sigma}}(ha) - \rho_1^{\pi^t}(ha)\right)\left(\rho_{-1}^{\pi^{\mu,\sigma}}(ha) d_1^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-1}^{\pi^t}(ha) d_1^{\pi^t,\sigma}(h,a)\right) \\ &\quad + \mu \sum_{h\in H}\sum_{a\in A(h)} \left(\rho_2^{\pi^{\mu,\sigma}}(ha) - \rho_2^{\pi^t}(ha)\right)\left(\rho_{-2}^{\pi^{\mu,\sigma}}(ha) d_2^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-2}^{\pi^t}(ha) d_2^{\pi^t,\sigma}(h,a)\right)\end{aligned}$$

We further decompose the summation over histories according to the acting player. By the definition of $d_i^{\pi,\sigma}$, we have $d_i^{\pi,\sigma}(h,a) = -f^{\pi,\sigma}(h,a)$ when $h \in H_i$, and $d_i^{\pi,\sigma}(h,a) = f^{\pi,\sigma}(h,a)$ when $h \in H_{3-i}$. Then, the right-hand side of the above inequality can be rewritten as

$$\begin{aligned}&-\mu \sum_{h\in H_1}\sum_{a\in A(h)} \left(\rho_1^{\pi^{\mu,\sigma}}(ha) - \rho_1^{\pi^t}(ha)\right)\left(\rho_{-1}^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-1}^{\pi^t}(ha) f^{\pi^t,\sigma}(h,a)\right) \\ &\quad + \mu \sum_{h\in H_2}\sum_{a\in A(h)} \left(\rho_1^{\pi^{\mu,\sigma}}(ha) - \rho_1^{\pi^t}(ha)\right)\left(\rho_{-1}^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-1}^{\pi^t}(ha) f^{\pi^t,\sigma}(h,a)\right) \\ &\quad - \mu \sum_{h\in H_2}\sum_{a\in A(h)} \left(\rho_2^{\pi^{\mu,\sigma}}(ha) - \rho_2^{\pi^t}(ha)\right)\left(\rho_{-2}^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-2}^{\pi^t}(ha) f^{\pi^t,\sigma}(h,a)\right) \\ &\quad + \mu \sum_{h\in H_1}\sum_{a\in A(h)} \left(\rho_2^{\pi^{\mu,\sigma}}(ha) - \rho_2^{\pi^t}(ha)\right)\left(\rho_{-2}^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho_{-2}^{\pi^t}(ha) f^{\pi^t,\sigma}(h,a)\right)\end{aligned}$$

Expanding the products in the preceding expression, we obtain

$$
\begin{aligned}
&- \mu \sum_{h \in H_1} \sum_{a \in A(h)} \Big(\rho^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^t,\sigma}(h,a) \\
&\qquad\qquad - \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) + \rho^{\pi^t}(ha) f^{\pi^t}(h,a) \Big) \\
&+ \mu \sum_{h \in H_2} \sum_{a \in A(h)} \Big(\rho^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^t,\sigma}(h,a) \\
&\qquad\qquad - \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) + \rho^{\pi^t}(ha) f^{\pi^t}(h,a) \Big) \\
&- \mu \sum_{h \in H_2} \sum_{a \in A(h)} \Big(\rho^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^t,\sigma}(h,a) \\
&\qquad\qquad - \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) + \rho^{\pi^t}(ha) f^{\pi^t}(h,a) \Big) \\
&+ \mu \sum_{h \in H_1} \sum_{a \in A(h)} \Big(\rho^{\pi^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^t,\sigma}(h,a) \\
&\qquad\qquad - \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) + \rho^{\pi^t}(ha) f^{\pi^t}(h,a) \Big)
\end{aligned}
$$

Collecting the terms over H_1 and H_2 separately, the terms with $\rho^{\pi^{\mu,\sigma}}(ha)$ and $\rho^{\pi^t}(ha)$ cancel out. Hence the preceding expression reduces to

$$
\begin{aligned}
&- \mu \sum_{h \in H_1} \sum_{a \in A(h)} \Big(\rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^t,\sigma}(h,a) \\
&\qquad\qquad + \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^t,\sigma}(h,a) - \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) \Big) \\
&- \mu \sum_{h \in H_2} \sum_{a \in A(h)} \Big(\rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) - \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) f^{\pi^t,\sigma}(h,a) \\
&\qquad\qquad + \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^t,\sigma}(h,a) - \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) f^{\pi^{\mu,\sigma},\sigma}(h,a) \Big) \\
= &-\mu \sum_{h \in H_1} \sum_{a \in A(h)} \left(\rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) \left(f^{\pi^{\mu,\sigma},\sigma}(h,a) - f^{\pi^t,\sigma}(h,a) \right) + \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) \left(f^{\pi^t,\sigma}(h,a) - f^{\pi^{\mu,\sigma},\sigma}(h,a) \right) \right) \\
&- \mu \sum_{h \in H_2} \sum_{a \in A(h)} \left(\rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha) \left(f^{\pi^{\mu,\sigma},\sigma}(h,a) - f^{\pi^t,\sigma}(h,a) \right) + \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha) \left(f^{\pi^t,\sigma}(h,a) - f^{\pi^{\mu,\sigma},\sigma}(h,a) \right) \right).
\end{aligned}
$$

Next, we substitute the explicit form of $f^{\pi,\sigma}$ into the preceding expression. To make the cancellation of the σ-dependent terms transparent, we first write

$$
\begin{aligned}
f^{\pi^{\mu,\sigma},\sigma}(h,a) &:= \frac{\partial \phi(\pi^{\mu,\sigma}(h))}{\partial \pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla \phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h) \rangle - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} - \phi(\sigma(h)) + \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \\
f^{\pi^t,\sigma}(h,a) &:= \frac{\partial \phi(\pi^t(h))}{\partial \pi(a|h)} + \phi(\pi^t(h)) - \langle \nabla \phi(\pi^t(h)), \pi^t(h) \rangle - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} - \phi(\sigma(h)) + \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle
\end{aligned}
$$

Substituting these expressions into the preceding bound gives

$$
\begin{aligned}
&\frac{d}{dt} J(t; \pi^{\mu,\sigma}) \\
&\leq -\mu \sum_{h \in H_1} \sum_{a \in A(h)} \rho^{\pi_1^{\mu,\sigma}, \pi_2^t}(ha) \cdot \\
&\quad \Bigg(\frac{\partial \phi(\pi^{\mu,\sigma}(h))}{\partial \pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla \phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h) \rangle - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} - \phi(\sigma(h)) + \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \\
&\qquad - \frac{\partial \phi(\pi^t(h))}{\partial \pi(a|h)} - \phi(\pi^t(h)) + \langle \nabla \phi(\pi^t(h)), \pi^t(h) \rangle + \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} + \phi(\sigma(h)) - \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \Bigg) \\
&\quad + \mu \sum_{h \in H_1} \sum_{a \in A(h)} \rho^{\pi_1^t, \pi_2^{\mu,\sigma}}(ha) \cdot \\
&\quad \Bigg(\frac{\partial \phi(\pi^{\mu,\sigma}(h))}{\partial \pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla \phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h) \rangle - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} - \phi(\sigma(h)) + \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \\
&\qquad - \frac{\partial \phi(\pi^t(h))}{\partial \pi(a|h)} - \phi(\pi^t(h)) + \langle \nabla \phi(\pi^t(h)), \pi^t(h) \rangle + \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} + \phi(\sigma(h)) - \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \Bigg) \\
&\quad - \mu \sum_{h \in H_2} \sum_{a \in A(h)} \rho^{\pi_1^t, \pi_2^{\mu,\sigma}}(ha) \cdot \\
&\quad \Bigg(\frac{\partial \phi(\pi^{\mu,\sigma}(h))}{\partial \pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla \phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h) \rangle - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} - \phi(\sigma(h)) + \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \\
&\qquad - \frac{\partial \phi(\pi^t(h))}{\partial \pi(a|h)} - \phi(\pi^t(h)) + \langle \nabla \phi(\pi^t(h)), \pi^t(h) \rangle + \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} + \phi(\sigma(h)) - \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \Bigg) \\
&\quad + \mu \sum_{h \in H_2} \sum_{a \in A(h)} \rho^{\pi_1^{\mu,\sigma}, \pi_2^t}(ha) \cdot \\
&\quad \Bigg(\frac{\partial \phi(\pi^{\mu,\sigma}(h))}{\partial \pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla \phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h) \rangle - \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} - \phi(\sigma(h)) + \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \\
&\qquad - \frac{\partial \phi(\pi^t(h))}{\partial \pi(a|h)} - \phi(\pi^t(h)) + \langle \nabla \phi(\pi^t(h)), \pi^t(h) \rangle + \frac{\partial \phi(\sigma(h))}{\partial \pi(a|h)} + \phi(\sigma(h)) - \langle \nabla \phi(\sigma(h)),\, \sigma(h) \rangle \Bigg)
\end{aligned}
$$

The σ-dependent terms cancel within each parenthesis. Therefore, the preceding bound simplifies to

$$
\begin{aligned}
&\frac{d}{dt}J(t;\pi^{\mu,\sigma}) \\
&\leq -\mu \sum_{h\in H_1}\sum_{a\in A(h)} \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha)\cdot \\
&\quad \left(\frac{\partial\phi(\pi^{\mu,\sigma}(h))}{\partial\pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle\nabla\phi(\pi^{\mu,\sigma}(h)),\pi^{\mu,\sigma}(h)\rangle - \frac{\partial\phi(\pi^t(h))}{\partial\pi(a|h)} - \phi(\pi^t(h)) + \langle\nabla\phi(\pi^t(h)),\pi^t(h)\rangle\right) \\
&\quad + \mu \sum_{h\in H_1}\sum_{a\in A(h)} \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha)\cdot \\
&\quad \left(\frac{\partial\phi(\pi^{\mu,\sigma}(h))}{\partial\pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle\nabla\phi(\pi^{\mu,\sigma}(h)),\pi^{\mu,\sigma}(h)\rangle - \frac{\partial\phi(\pi^t(h))}{\partial\pi(a|h)} - \phi(\pi^t(h)) + \langle\nabla\phi(\pi^t(h)),\pi^t(h)\rangle\right) \\
&\quad - \mu \sum_{h\in H_2}\sum_{a\in A(h)} \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(ha)\cdot \\
&\quad \left(\frac{\partial\phi(\pi^{\mu,\sigma}(h))}{\partial\pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle\nabla\phi(\pi^{\mu,\sigma}(h)),\pi^{\mu,\sigma}(h)\rangle - \frac{\partial\phi(\pi^t(h))}{\partial\pi(a|h)} - \phi(\pi^t(h)) + \langle\nabla\phi(\pi^t(h)),\pi^t(h)\rangle\right) \\
&\quad + \mu \sum_{h\in H_2}\sum_{a\in A(h)} \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(ha)\cdot \\
&\quad \left(\frac{\partial\phi(\pi^{\mu,\sigma}(h))}{\partial\pi(a|h)} + \phi(\pi^{\mu,\sigma}(h)) - \langle\nabla\phi(\pi^{\mu,\sigma}(h)),\pi^{\mu,\sigma}(h)\rangle - \frac{\partial\phi(\pi^t(h))}{\partial\pi(a|h)} - \phi(\pi^t(h)) + \langle\nabla\phi(\pi^t(h)),\pi^t(h)\rangle\right)
\end{aligned}
$$

Using $\rho^\pi(ha) = \rho^\pi(h)\pi(a|h)$ and $\sum_{a\in A(h)} \pi(a|h) = 1$, we collect the sums over actions into inner products. This yields

$$
\begin{aligned}
&\frac{d}{dt} J(t;\pi^{\mu,\sigma}) \\
&\leq -\mu \sum_{h\in H_1} \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(h) \left(\langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h)\rangle + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h)\rangle \right. \\
&\qquad\qquad \left. - \langle \nabla\phi(\pi^t(h)), \pi^{\mu,\sigma}\rangle - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^t(h)), \pi^t(h)\rangle \right) \\
&\quad + \mu \sum_{h\in H_1} \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(h) \left(\langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^t\rangle + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h)\rangle \right. \\
&\qquad\qquad \left. - \langle \nabla\phi(\pi^t(h)), \pi^t\rangle - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^t(h)), \pi^t(h)\rangle \right) \\
&\quad - \mu \sum_{h\in H_2} \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(h) \left(\langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h)\rangle + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h)\rangle \right. \\
&\qquad\qquad \left. - \langle \nabla\phi(\pi^t(h)), \pi^{\mu,\sigma}\rangle - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^t(h)), \pi^t(h)\rangle \right) \\
&\quad + \mu \sum_{h\in H_2} \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(h) \left(\langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^t\rangle + \phi(\pi^{\mu,\sigma}(h)) - \langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^{\mu,\sigma}(h)\rangle \right. \\
&\qquad\qquad \left. - \langle \nabla\phi(\pi^t(h)), \pi^t\rangle - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^t(h)), \pi^t(h)\rangle \right) \\
&= -\mu \sum_{h\in H_1} \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(h) \left(\phi(\pi^{\mu,\sigma}(h)) - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^t(h)), \pi^t(h) - \pi^{\mu,\sigma}(h)\rangle \right) \\
&\quad + \mu \sum_{h\in H_1} \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(h) \left(\phi(\pi^{\mu,\sigma}(h)) - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^t(h) - \pi^{\mu,\sigma}(h)\rangle \right) \\
&\quad - \mu \sum_{h\in H_2} \rho^{\pi_1^t,\pi_2^{\mu,\sigma}}(h) \left(\phi(\pi^{\mu,\sigma}(h)) - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^t(h)), \pi^t(h) - \pi^{\mu,\sigma}(h)\rangle \right) \\
&\quad + \mu \sum_{h\in H_2} \rho^{\pi_1^{\mu,\sigma},\pi_2^t}(h) \left(\phi(\pi^{\mu,\sigma}(h)) - \phi(\pi^t(h)) + \langle \nabla\phi(\pi^{\mu,\sigma}(h)), \pi^t(h) - \pi^{\mu,\sigma}(h)\rangle \right) \\
&= -\mu \sum_{i\in N} \sum_{h\in H_i} \left(\rho^{\pi_i^{\mu,\sigma},\pi_{-i}^t}(h) D_\phi(\pi^{\mu,\sigma}(h), \pi^t(h)) + \rho^{\pi_i^t,\pi_{-i}^{\mu,\sigma}}(h) D_\phi(\pi^t(h), \pi^{\mu,\sigma}(h)) \right).
\end{aligned}
$$

□

H Proof of Theorem 9

H.1 Proof of Lemma 10

Proof.

$$
\sum_{i\in N} \left(v_i^{\pi_i^*,\sigma_{-i}^{k+1}}(h_{\text{init}}) - v_i^{\sigma^{k+1}}(h_{\text{init}}) \right) - \mu \sum_{i\in N} \sum_{x\in\mathcal{X}_i} \sum_{a\in A(x)} \left(\rho_i^{\pi^*}(x)\pi^*(a|x) - \rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x) \right) \cdot \left(\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)} - \frac{\partial\phi(\sigma^k(x))}{\partial\pi(a|x)} + D_\phi(\sigma^{k+1}(x), \sigma^k(x)) - \langle \nabla\phi(\sigma^{k+1}(x)) - \nabla\phi(\sigma^k(x)), \sigma^{k+1}(x)\rangle \right) \leq 0.
$$

Thus,

$$
\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)\cdot\\
&\quad\left(\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)}-\frac{\partial\phi(\sigma^{k}(x))}{\partial\pi(a|x)}+D_\phi(\sigma^{k+1}(x),\sigma^k(x))-\langle\nabla\phi(\sigma^{k+1}(x))-\nabla\phi(\sigma^k(x)),\sigma^{k+1}(x)\rangle\right)\\
&\leq\frac{1}{\mu}\sum_{i\in N}\left(v_i^{\sigma^{k+1}}(h_{\text{init}})-v_i^{\pi_i^*,\sigma_{-i}^{k+1}}(h_{\text{init}})\right)\\
&\leq\frac{1}{\mu}\sum_{i\in N}\left(v_i^{\pi_i^*,\sigma_{-i}^{k+1}}(h_{\text{init}})-v_i^{\pi^*}(h_{\text{init}})\right)\\
&\leq 0,
\end{aligned}
$$

where the second inequality stems from the monotonicity assumption. Hence, we have:

$$
\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)\cdot\\
&\quad\left(\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)}-\frac{\partial\phi(\sigma^{k}(x))}{\partial\pi(a|x)}+D_\phi(\sigma^{k+1}(x),\sigma^k(x))-\langle\nabla\phi(\sigma^{k+1}(x))-\nabla\phi(\sigma^k(x)),\sigma^{k+1}(x)\rangle\right)\leq 0.
\end{aligned}
$$

Here,

$$
\begin{aligned}
&\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)D_\phi(\pi^*(x),\sigma^{k+1}(x))-\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)D_\phi(\pi^*(x),\sigma^{k}(x))\\
&=\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)\left(\phi(\pi^*(x))-\phi(\sigma^{k+1}(x))-\langle\nabla\phi(\sigma^{k+1}(x)),\pi^*(x)-\sigma^{k+1}(x)\rangle\right)\\
&\quad-\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)\left(\phi(\pi^*(x))-\phi(\sigma^{k}(x))-\langle\nabla\phi(\sigma^{k}(x)),\pi^*(x)-\sigma^{k}(x)\rangle\right)\\
&=\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)\left(-\phi(\sigma^{k+1}(x))+\phi(\sigma^{k}(x))-\langle\nabla\phi(\sigma^{k+1}(x)),\pi^*(x)-\sigma^{k+1}(x)\rangle+\langle\nabla\phi(\sigma^{k}(x)),\pi^*(x)-\sigma^{k}(x)\rangle\right)\\
&=\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)\left(-D_\phi(\sigma^{k+1}(x),\sigma^{k}(x))-\langle\nabla\phi(\sigma^{k+1}(x)),\pi^*(x)-\sigma^{k+1}(x)\rangle+\langle\nabla\phi(\sigma^{k}(x)),\pi^*(x)-\sigma^{k+1}(x)\rangle\right)\\
&=-\sum_{x\in\mathcal{X}_i}\rho_i^{\sigma^{k+1}}(x)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x))+\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\pi^*}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)}\\
&\quad-\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\pi^*}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)\frac{\partial\phi(\sigma^{k}(x))}{\partial\pi(a|x)}\\
&\quad+\sum_{x\in\mathcal{X}_i}\left(\rho_i^{\sigma^{k+1}}(x)-\rho_i^{\pi^*}(x)\right)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x))\\
&=-\sum_{x\in\mathcal{X}_i}\rho_i^{\sigma^{k+1}}(x)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x))+\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)}\\
&\quad-\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)\frac{\partial\phi(\sigma^{k}(x))}{\partial\pi(a|x)}\\
&\quad-\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\sigma^{k+1}(a|x)\right)\left(\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)}-\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)}\right)\\
&\quad+\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x))\\
&=-\sum_{x\in\mathcal{X}_i}\rho_i^{\sigma^{k+1}}(x)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x))+\sum_{x\in\mathcal{X}_i}\sum_{a\in A(x)}\left(\rho_i^{\sigma^{k+1}}(x)\sigma^{k+1}(a|x)-\rho_i^{\pi^*}(x)\pi^*(a|x)\right)\cdot\\
&\left(\frac{\partial\phi(\sigma^{k+1}(x))}{\partial\pi(a|x)}-\frac{\partial\phi(\sigma^{k}(x))}{\partial\pi(a|x)}+D_\phi(\sigma^{k+1}(x),\sigma^{k}(x))-\langle\nabla\phi(\sigma^{k+1}(x))-\nabla\phi(\sigma^{k}(x)),\sigma^{k+1}(x)\rangle\right).
\end{aligned}
$$

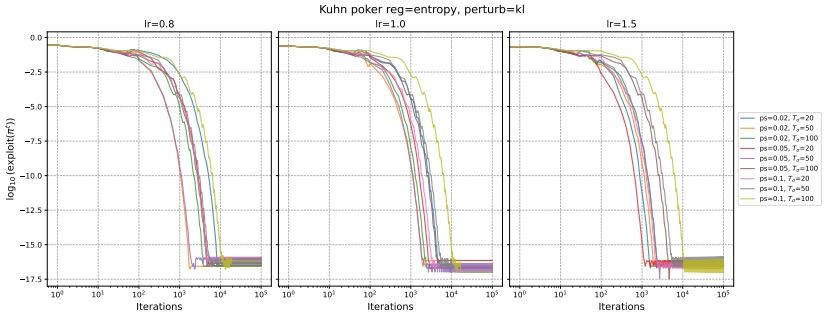
Combining these inequalities, we obtain:

$$
\begin{aligned}
&\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)D_\phi(\pi^*(x),\sigma^{k+1}(x))-\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\pi^*}(x)D_\phi(\pi^*(x),\sigma^{k}(x))\\
&\le-\sum_{i\in N}\sum_{x\in\mathcal{X}_i}\rho_i^{\sigma^{k+1}}(x)D_\phi(\sigma^{k+1}(x),\sigma^{k}(x)).
\end{aligned}
$$

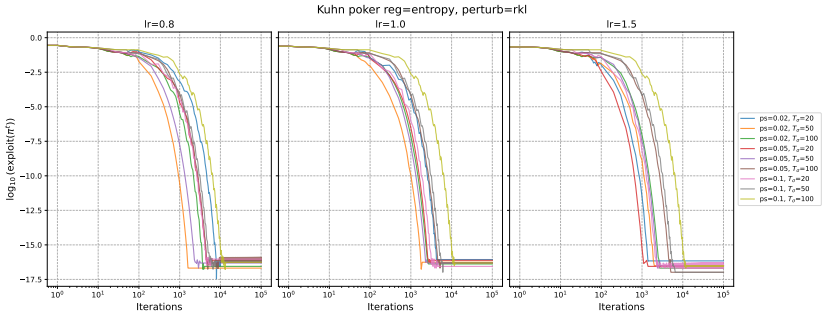
□

I Hyperparameter Tuning Results

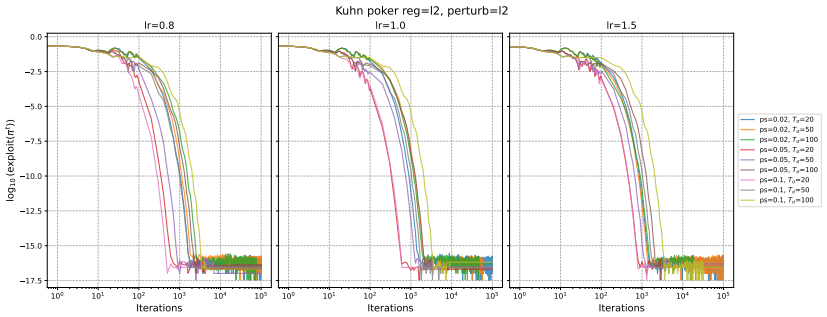
This appendix reports the hyperparameter tuning results used for the experimental comparisons in Section 6. For PFTRL and 0s-PFTRL, we sweep the learning rate η, the anchoring strategy update frequency T_σ, and the perturbation strength μ under three perturbation-regularizer combinations: entropy regularization with KL perturbation, entropy regularization with reverse KL perturbation, and squared-ℓ_2 regularization with squared-ℓ_2 perturbation. For PCFR+ and 0s-PCFR+, we sweep the anchoring strategy update frequency T_σ and the perturbation strength μ using squared-ℓ_2 perturbation. In the legends, "ps" denotes the perturbation strength μ.



(a) PFTRL-KL



(b) PFTRL-RKL



(c) PFTRL-L2

Figure 2: PFTRL hyperparameter tuning results on Kuhn poker.

I.1 PFTRL

We first present the tuning results for PFTRL. The plots below compare the convergence behavior across the tested hyperparameter grid for each benchmark game. Unless otherwise stated, the configuration selected for the main experiments is the one that converges fastest among the tested settings; if no setting converges, we select the one with the lowest final exploitability.

I.1.1 Kuhn Poker

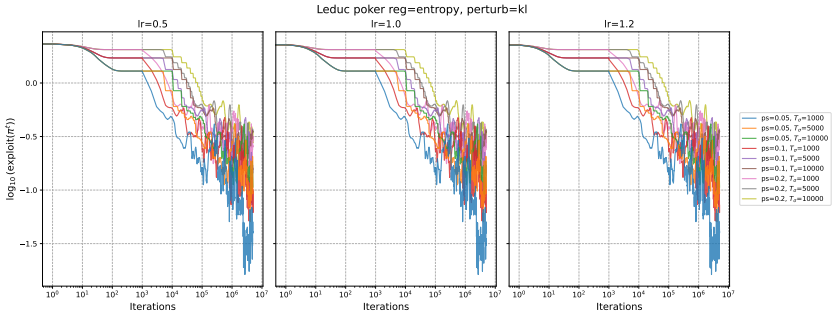
Figure 2 shows the PFTRL hyperparameter tuning results on Kuhn poker.

I.1.2 Leduc Poker

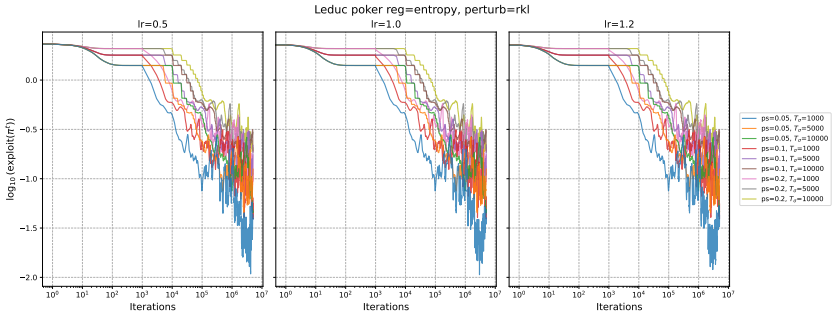
Figure 3 shows the PFTRL hyperparameter tuning results on Leduc poker.

I.1.3 Goofspiel (4 Cards)

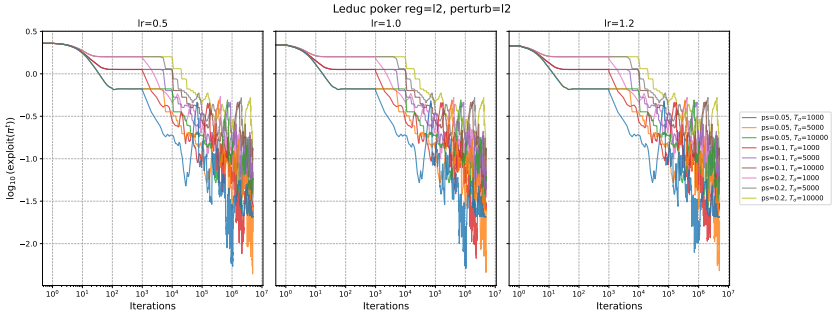
Figure 4 reports the PFTRL hyperparameter tuning results on Goofspiel with 4 cards.



(a) PFTRL-KL



(b) PFTRL-RKL



(c) PFTRL-L2

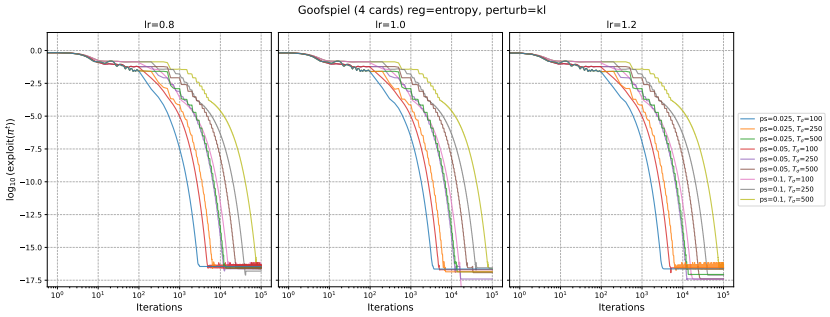
Figure 3: PFTRL hyperparameter tuning results on Leduc poker.

I.1.4 Goofspiel (5 Cards)

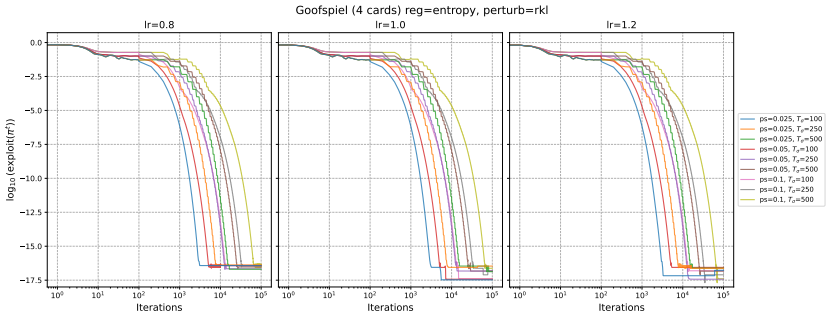
Figure 5 reports the PFTRL hyperparameter tuning results on Goofspiel with 5 cards.

I.1.5 Liar's Dice (4 Dice Sides)

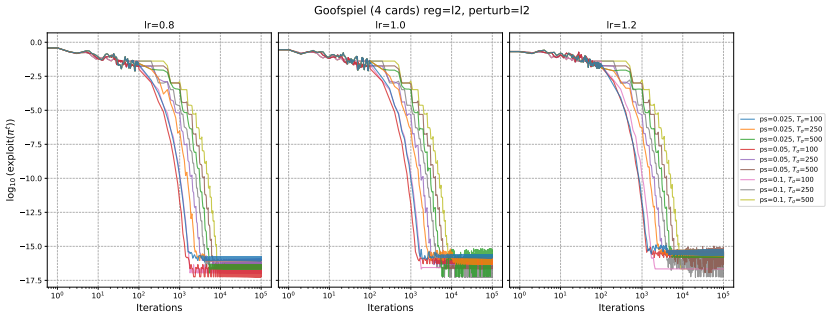
Figure 6 reports the PFTRL hyperparameter tuning results on Liar's Dice with 4 dice sides. We also conducted additional experiments on the same game; the results are shown in Figure 7.



(a) PFTRL-KL

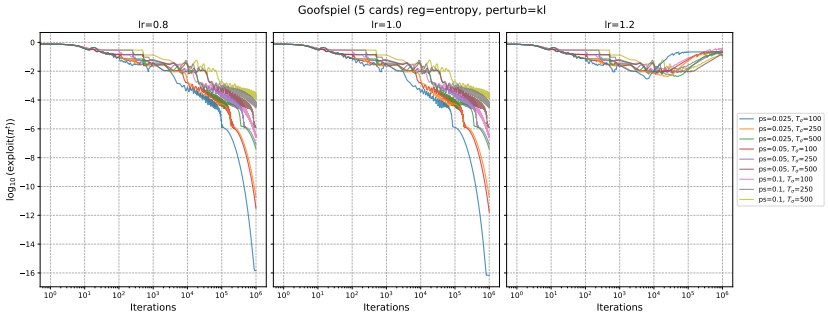


(b) PFTRL-RKL

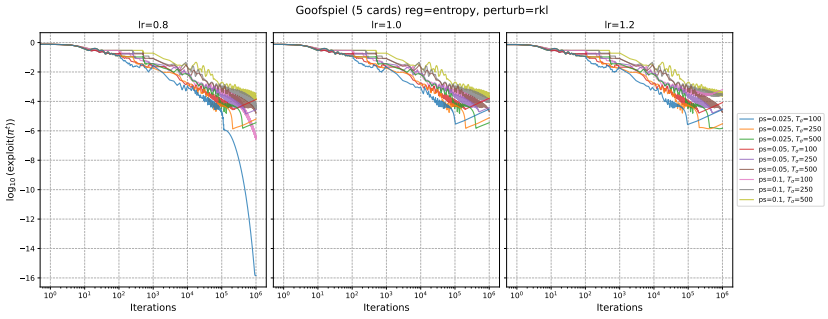


(c) PFTRL-L2

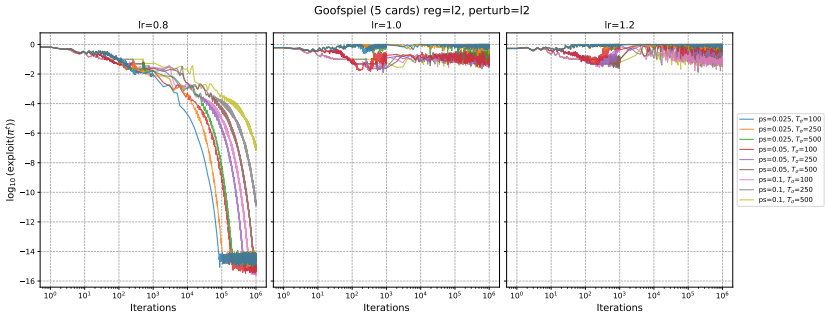
Figure 4: PFTRL hyperparameter tuning results on Goofspiel with 4 cards.



(a) PFTRL-KL

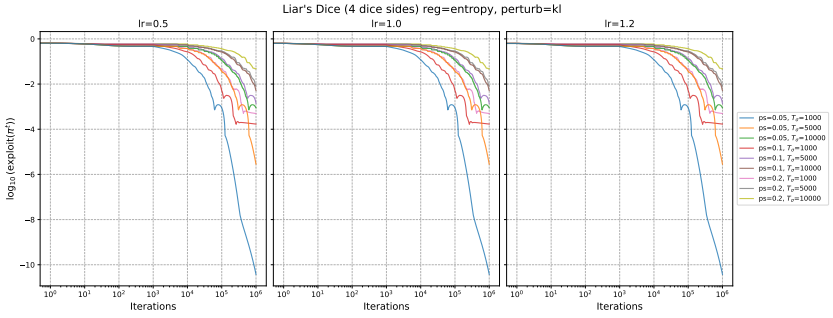


(b) PFTRL-RKL

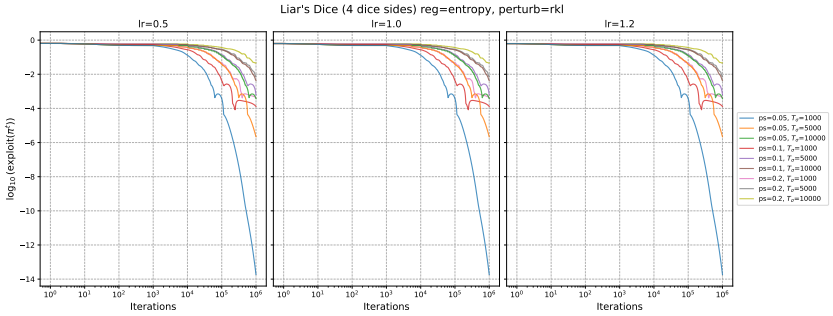


(c) PFTRL-L2

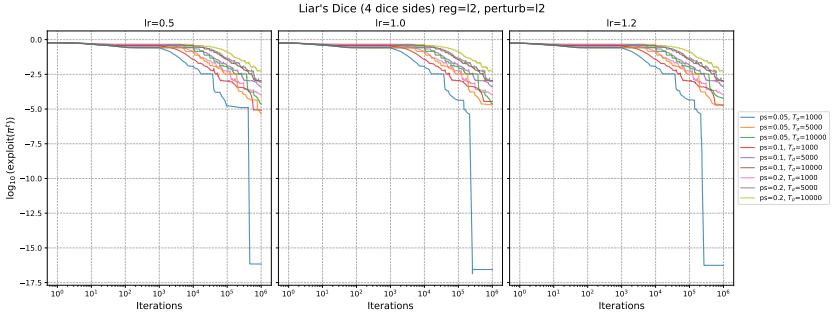
Figure 5: PFTRL hyperparameter tuning results on Goofspiel with 5 cards.



(a) PFTRL-KL

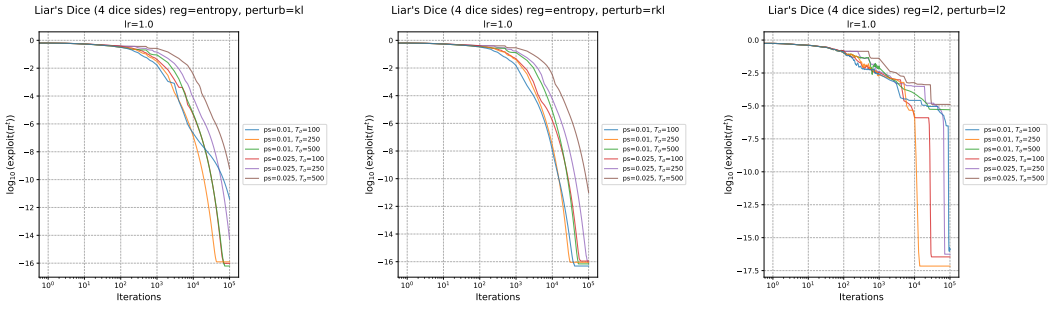


(b) PFTRL-RKL



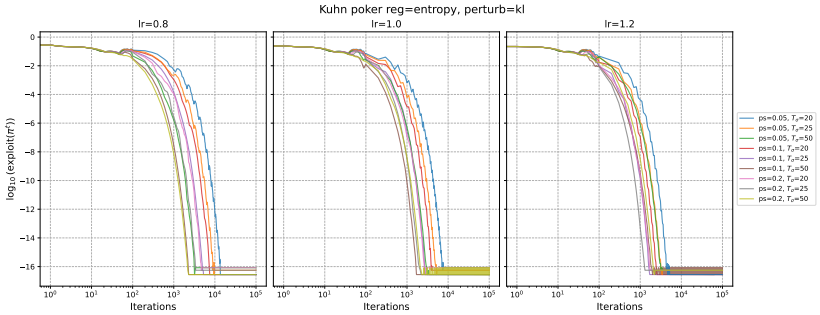
(c) PFTRL-L2

Figure 6: PFTRL hyperparameter tuning results on Liar's Dice with 4 dice sides.

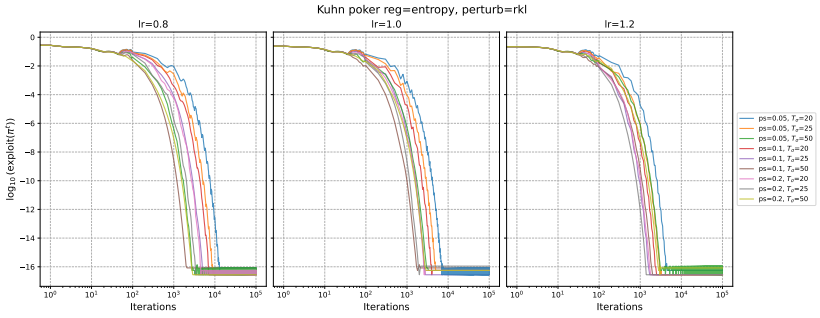


(a) PFTRL-KL (b) PFTRL-RKL (c) PFTRL-L2

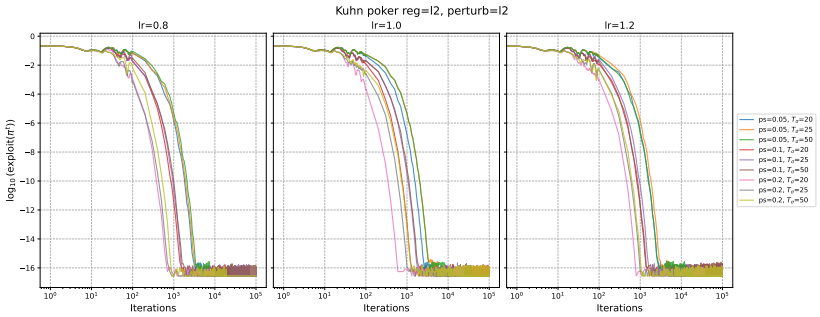
Figure 7: Additional PFTRL hyperparameter tuning results on Liar's Dice with 4 dice sides.



(a) 0s-PFTRL-KL



(b) 0s-PFTRL-RKL



(c) 0s-PFTRL-L2

Figure 8: 0s-PFTRL hyperparameter tuning results on Kuhn poker.

I.2 0s-PFTRL

We report the corresponding tuning results for 0s-PFTRL.

I.2.1 Kuhn Poker

Figure 8 shows the 0s-PFTRL hyperparameter tuning results on Kuhn poker.

I.2.2 Leduc Poker

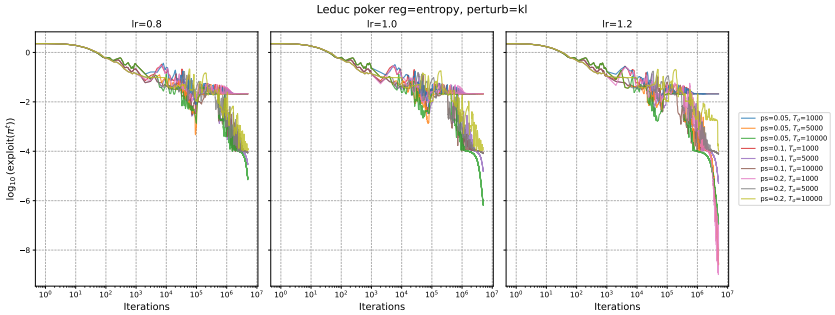
Figure 9 reports the 0s-PFTRL hyperparameter tuning results on Leduc poker.

I.2.3 Goofspiel (4 Cards)

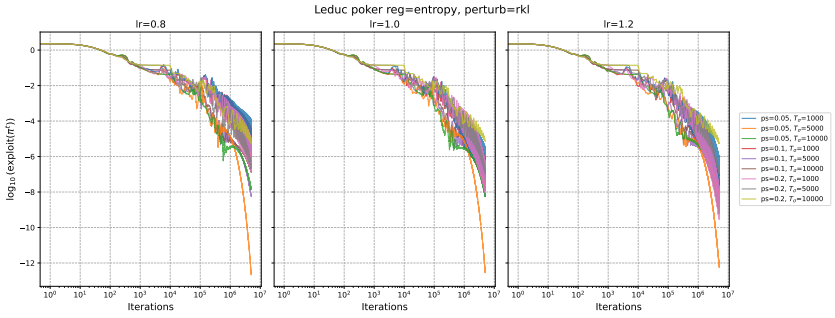
Figure 10 reports the 0s-PFTRL hyperparameter tuning results on Goofspiel with 4 cards.

I.2.4 Goofspiel (5 Cards)

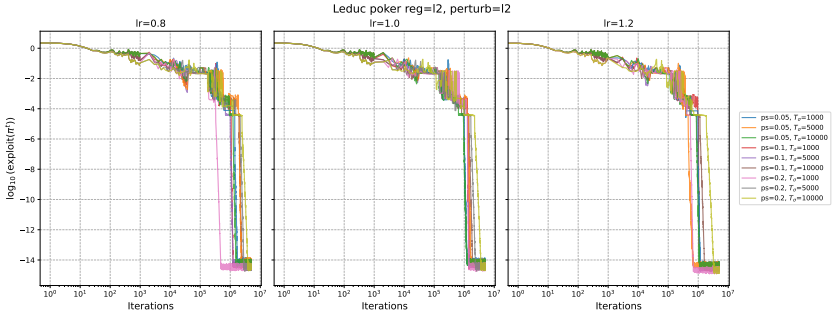
Figure 11 reports the 0s-PFTRL hyperparameter tuning results on Goofspiel with 5 cards.



(a) 0s-PFTRL-KL



(b) 0s-PFTRL-RKL

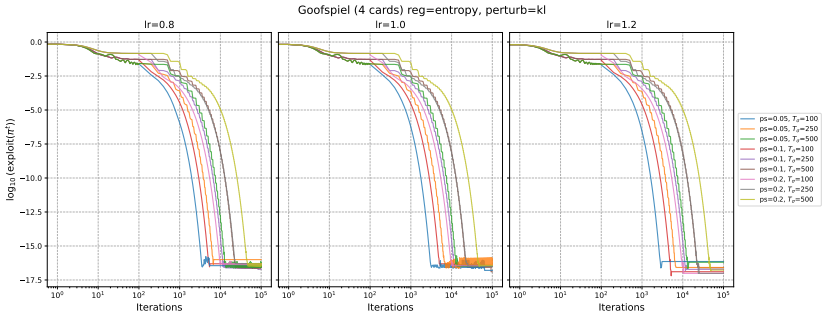


(c) 0s-PFTRL-L2

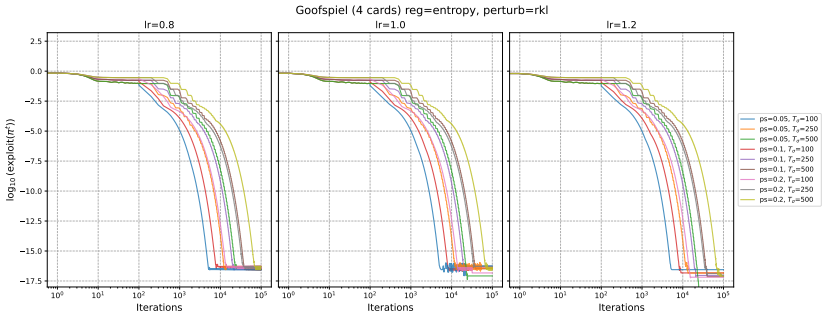
Figure 9: 0s-PFTRL hyperparameter tuning results on Leduc poker.

I.2.5 Liar's Dice (4 Dice Sides)

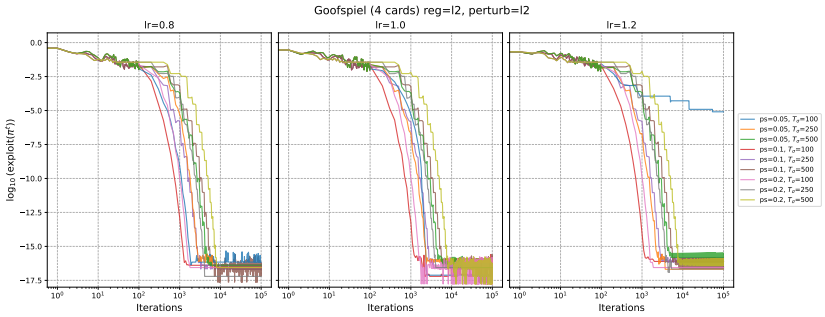
Figure 12 reports the 0s-PFTRL hyperparameter tuning results on Liar's Dice with 4 dice sides.



(a) 0s-PFTRL-KL

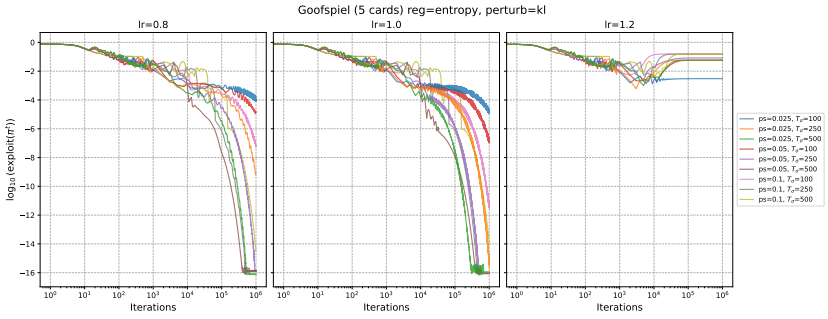


(b) 0s-PFTRL-RKL

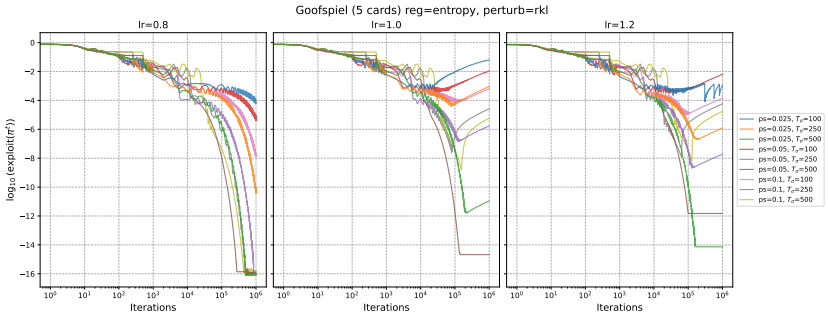


(c) 0s-PFTRL-L2

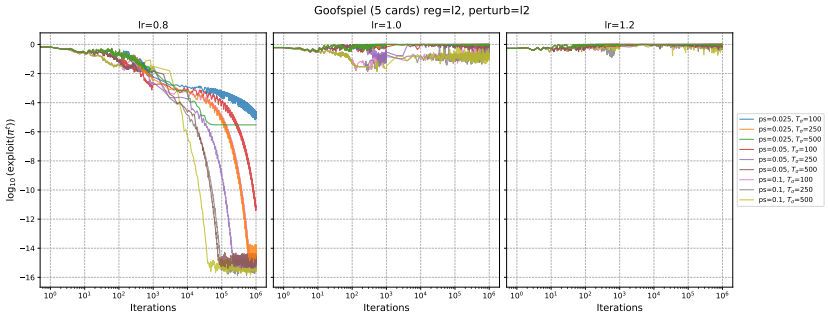
Figure 10: 0s-PFTRL hyperparameter tuning results on Goofspiel with 4 cards.



(a) 0s-PFTRL-KL

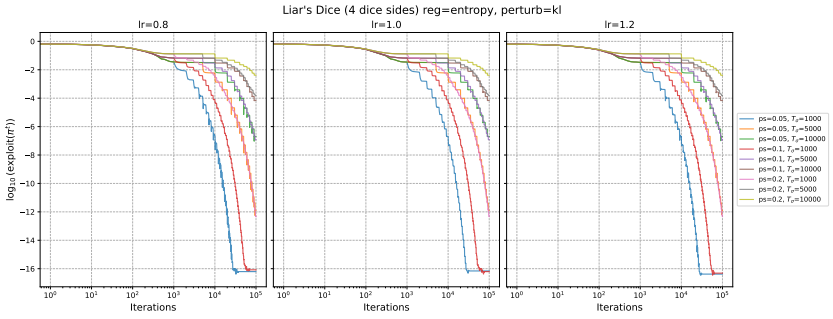


(b) 0s-PFTRL-RKL

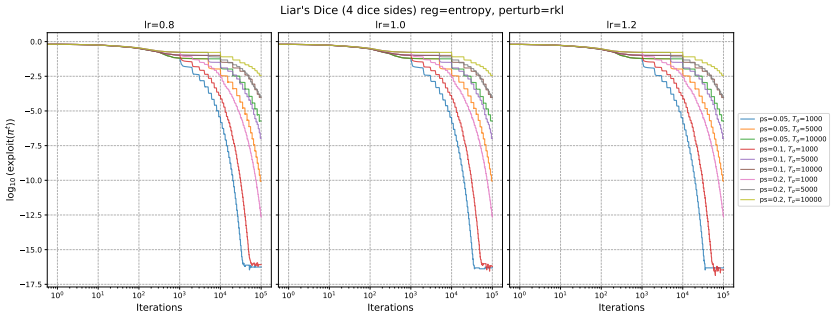


(c) 0s-PFTRL-L2

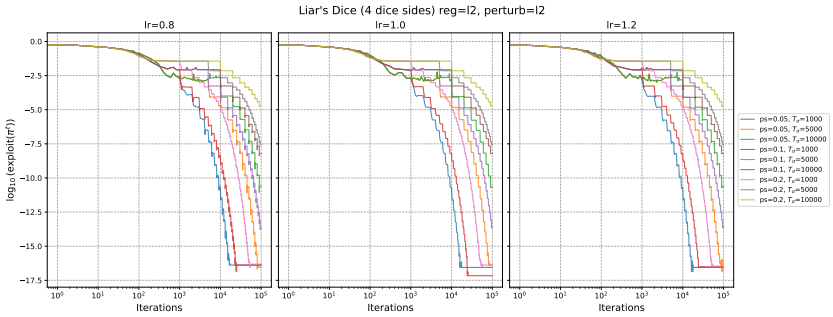
Figure 11: 0s-PFTRL hyperparameter tuning results on Goofspiel with 5 cards.



(a) 0s-PFTRL-KL

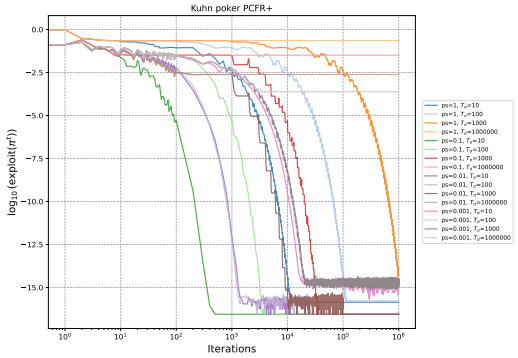


(b) 0s-PFTRL-RKL

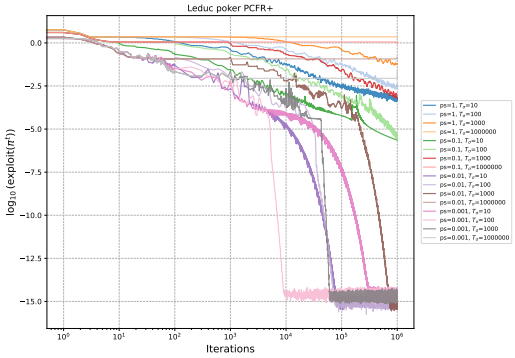


(c) 0s-PFTRL-L2

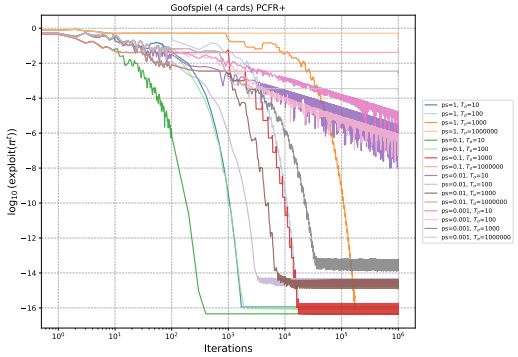
Figure 12: 0s-PFTRL hyperparameter tuning results on Liar's Dice with 4 dice sides.



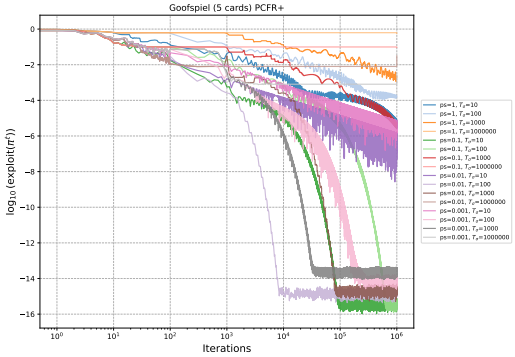
(a) Kuhn poker



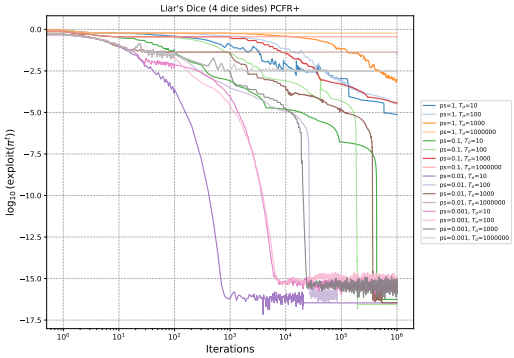
(b) Leduc poker



(c) Goofspiel with 4 cards



(d) Goofspiel with 5 cards



(e) Liar's Dice with 4 dice sides

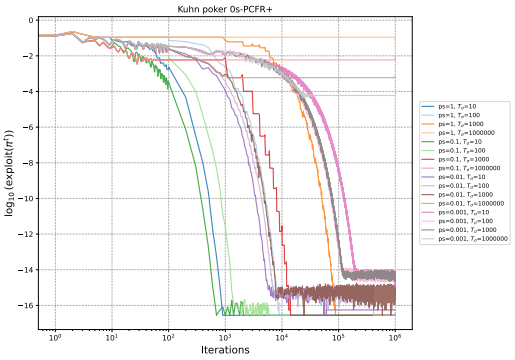
Figure 13: PCFR+ hyperparameter tuning results on each game.

I.3 PCFR+

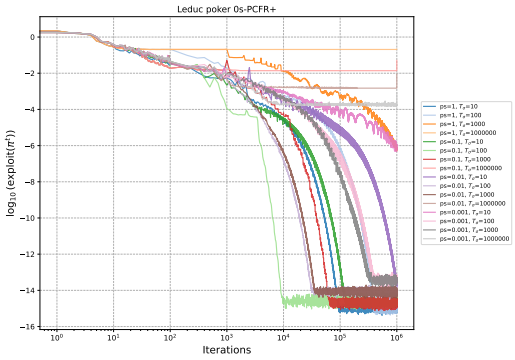
Figure 13 reports the PCFR+ hyperparameter tuning results on each game.

I.4 0s-PCFR+

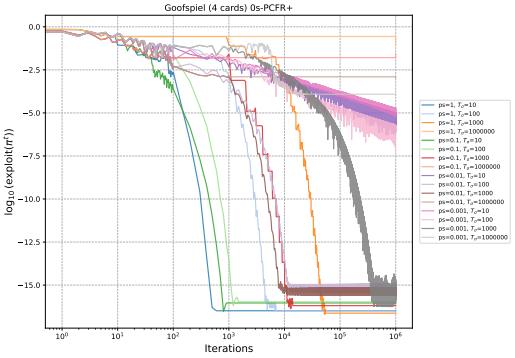
Figure 14 reports the 0s-PCFR+ hyperparameter tuning results on each game.



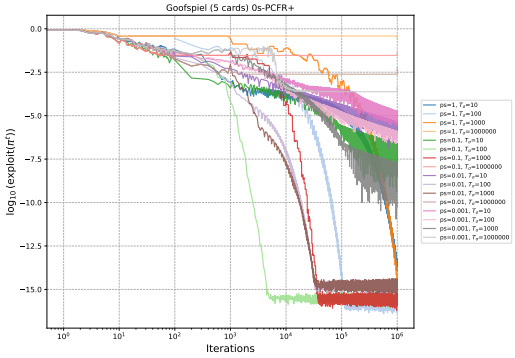
(a) Kuhn poker



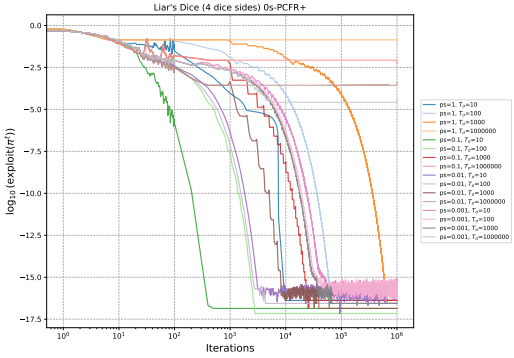
(b) Leduc poker



(c) Goofspiel with 4 cards



(d) Goofspiel with 5 cards



(e) Liar's Dice with 4 dice sides

Figure 14: 0s-PCFR+ hyperparameter tuning results on each game.