Fast Computation of Nash Equilibria in Imperfect Information Games

Remi Munos [1] **Julien Perolat** [1] **Jean-Baptiste Lespiau** [1] **Mark Rowland** [1] **Bart De Vylder** [1] **Marc Lanctot** [1] **Finbarr Timbers** [1] **Daniel Hennes** [1] **Shayegan Omidshafiei** [1] **Audrunas Gruslys** [1] **Mohammad Gheshlaghi Azar** [1] **Edward Lockhart** [1] **Karl Tuyls** [1]

Abstract

We introduce and analyze a class of algorithms, called *Mirror Ascent against an Improved Opponent* (MAIO), for computing Nash equilibria in two-player zero-sum games, both in normal form and in sequential form with imperfect information. These algorithms update the policy of each player with a mirror-ascent step to maximize the value of playing against an improved opponent. An improved opponent can be a best response, a greedy policy, a policy improved by policy gradient, or by any other reinforcement learning or search techniques. We establish a convergence result of the last iterate to the set of Nash equilibria and show that the speed of convergence depends on the amount of improvement offered by these improved policies. In addition, we show that under some condition, if we use a best response as improved policy, then an exponential convergence rate is achieved.

1. Introduction

This paper considers the problem of computing a Nash equilibrium for two-player zero-sum games in two types of games: normal-form games and imperfect information games (IIGs) in extensive form. We introduce and analyze a class of algorithms, called **Mirror Ascent against an Improved Opponent (MAIO)**, which updates the policy of each player by following a step of mirror-ascent for maximizing its expected reward against an improved policy for the opponent. The actual implementation of the algorithm depends on how we choose to define the 'improved policy'.

If we use the best response (BR) (the opponent's best policy against the current player) as improved policy we show that, under some condition, the algorithm (MAIO-BR) produces a sequence of policies that converges to the set of Nash equilibria at an exponential rate. By that we mean that some weighted ℓ_2 distance between the policies produced by the algorithm and the set of Nash equilibria decreases as $O(\exp(-\beta t))$, for some problem-dependent constant $\beta > 0$, where t is the number of iterations of the algorithm.

However, in large IIGs it may be computationally prohibitive to compute a full best response at every iteration (since this is equivalent to solving an optimal control problem). Our analysis shows that the speed of convergence to the set of Nash equilibria depends on a measure (called the *improvement*) of how much each player is able to improve its own policy against a fixed opponent. In principle, the best response provides the best possible improvement, but due to its high computational cost other less-computationally expensive strategies can provide an improved policy as well at a lower computational cost. Examples of improved polices are the greedy policy (one-step policy improvement), a multi-step improved policy, such as in Monte Carlo Tree Search (MCTS), a policy improved by policy gradient, or by any other reinforcement learning or search algorithm. Our analysis shows convergence for all such cases, which opens new avenues for designing algorithms with convergence guarantees, while offering a trade-off in terms of computational cost versus convergence speed toward the Nash equilibrium.

Literature context: This work sits in the context of computing Nash equilibria for sequential games. One can distinguish several approaches to find a Nash equilibrium: (i) **Linear programming** has been the first approach applied to compute minimax equilibrium in imperfect information games (Von Stengel, 1996; Koller et al., 1994; Koller and Pfeffer, 1997) using sequence-form reductions methods. But these methods remain quite inefficient as the size of the action space grows even if linear programming methods (Khachiyan, 1980; Karmarkar, 1984; Nesterov and Todd, 1998) achieve exponential convergence to Nash equilibria in value, with a rate independent of game-dependent quantities. (ii) **Fictitious play** (FP) has been considered in the tabular case in (Heinrich et al., 2015) and with function approximation (Heinrich and Silver, 2016). In the normal

[1]DeepMind. Correspondence to: Remi Munos <munos@google.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

form case, FP has a proven convergence speed to the Nash equilibrium of $O(t^{\frac{-1}{m+n-2}})$, where m and n are the number of actions of each players. (iii) **Non-smooth convex optimization** has been one of the techniques providing the fastest rates of convergence (Nesterov, 2005; Hoda et al., 2010). In imperfect information games, one can achieve a rate of convergence of $O(\frac{1}{t})$ (Gilpin et al., 2007; Kroer et al., 2018) with an appropriate smoothing and an exponential convergence $O(\exp(-\kappa t))$ with a problem dependent constant κ (Gilpin et al., 2012) similar to ours. In terms of computational complexity, (Gilpin et al., 2012) requires processing a ℓ_2-projection onto the global space of strategies (sets of realization plans, also called treeplex). They propose an iterated algorithm to perform this projection, but this step is computationally much more involved than simple projections onto the simplex. On the contrary, our algorithm updates the policy at each state individually (e.g. for ℓ_2 regularization, we do a simple ℓ_2-projection onto the simplex at each state), which has a much lower computational complexity per iteration than projecting onto the treeplex. (iv) **Extragradient or optimistic mirror descent** methods have been proven to converge to a Nash equilibrium (Korpelevich, 1976) with possibly an exponential rate in unconstrained spaces (Mokhtari et al., 2020) but is not (to the best of our knowledge) applied in sequential form games. Furthermore, the analysis of extragradient methods is mostly done in unconstrained domains whereas the constrained domain considered here (where the constraints are the space of stochastic policies) remain more involved. The most closely related extragradient method in this domain is Optimistic Multiplicative-Weights-Update (OMWU) (Daskalakis and Panageas, 2018) which provides convergence guarantees to the Nash equilibrium of the last iterate (whilst most of the literature shows convergence of the average strategy (Daskalakis et al., 2011; Rakhlin and Sridharan, 2013; Kangarshahi et al., 2018) at a rate of $O(1/t)$). In Daskalakis and Panageas (2018), the authors conjecture that this technique can be useful to prove the convergence of the last iterate of many algorithms. Our analysis generalizes this approach beyond OMWU and beyond normal-form games. A related approach uses the Frank-Wolfe method to compute Nash equilibria in normal-form games (Gidel et al., 2016), although convergence is attained at the same rate as for fictitious play. (v) **Regret minimization** has been extensively considered in games since the average strategy of self-playing no-regret algorithms converges to a Nash equilibrium (Rakhlin and Sridharan, 2013; Kangarshahi et al., 2018) and provides a fast rate of $O(\frac{1}{t})$ (Syrgkanis et al., 2015). This technique is usually studied in the discrete time setting but has also been looked at in continuous time (Mertikopoulos et al., 2018). Finally, the main state-of-the-art methods in IIGs remain **counterfactual regret minimization** (CFR) (Zinkevich et al., 2008) and has been studied extensively in zero-sum imperfect information games. In its most simple form all players learn in self-play to update their strategy at each information state according to a regret minimizing algorithm on the counterfactual value of the joint policy. In that setting the average policy played by all players converges to a Nash equilibrium with a $O(1/\sqrt{t})$ rate. The standard method has seen many improvements (for example the CFR+ algorithm of (Tammelin et al., 2015)). The convergence of an iterate (not necessarily the last) can be achieved if players use a regret minimization strategy against a best responding opponent (Johanson et al., 2012; Lockhart et al., 2019) in time $O(1/(p\sqrt{t}))$.

Our contribution: This work sits at the intersection of counterfactual regret-minimization and extragradient approaches. We prove that the last iterate of MAIO converges to the set of Nash equilibria at a rate which depends on how much we are able to compute an improved policy at each step. When the improved policy is the best response, we achieve an exponential convergence (under some condition).

We also show convergence when the improved policy is the result of an extra-gradient step, or simply a greedy policy (much cheaper to compute than a best response) or a multi-step improvement such as implemented by a MCTS algorithm. This sheds a new light on the relation between seemingly different approaches (e.g., CFR-BR and extragradient methods) and proposes a whole spectrum of methods based on improvements.

Outline: We start by introducing MAIO in the normal-form game setting, then derive several variants depending on the type of regularization that is used (entropy or ℓ_2). An exponential convergence rate is achieved when using the best response as improved opponent (MAIO-BR). Subsequently, we consider the IIG setting, reporting convergence results and discussing the trade-off between (i) the computational complexity of finding an improved opponent and (ii) the speed of convergence toward the Nash equilibria. Finally, Section 4 reports numerical experiments on matrix games. The appendix contains all proofs as well as additional numerical experiments on IIGs.

2. Normal form games

In this section we consider the setting of games in normal form. The two players are indexed by $i \in \{1, 2\}$. A policy profile π refers to the set of policies used by each player $\pi = \{\pi_1, \pi_2\}$, where each policy $\pi_i \in \Delta(A_i)$ is a distribution over actions A_i available to player i. For simplicity we will omit the player index when it is obvious from the subscript, and use parentheses instead of braces, writing $\pi = (\pi_1, \pi_2) = (\pi_2, \pi_1)$. We will denote by A a generic action space when the reference to a specific player is not important.

The value of a policy profile $\pi = (\pi_1, \pi_2)$ is $V^\pi \stackrel{\text{def}}{=} \pi_1^\top R \pi_2$, where R is the payoff matrix of the game. Player 1 is trying to maximize the value whereas player 2 intends to minimize it. From the minimax theorem (Neumann, 1928), the (minimax) value of the game is

$$V^* \stackrel{\text{def}}{=} \max_{\pi_1} \min_{\pi_2} V^{(\pi_1,\pi_2)} = \min_{\pi_2} \max_{\pi_1} V^{(\pi_1,\pi_2)}$$

and is achieved for any $\pi \in \Pi^*$, where Π^* is the set of Nash equilibria of the game.

Additional notations: We write $V_1^\pi = V^\pi$ and $V_2^\pi = -V^\pi$, so each player $i \in \{1,2\}$ is trying to maximize (over π_i) the value V_i^π. We write $Q_i^{\pi_{-i}}$ for the payoff vector of player i against the opponent's policy π_{-i} (where $-i$ denotes player i's opponent). Thus $Q_1^{\pi_2} = R\pi_2$ and $Q_2^{\pi_1} = -R^\top \pi_1$. Notice that this notation will be further extended to a state-action Q-value function in the section on IIG.

The MAIO algorithm (defined below) will make use of the notion of an 'improved' policy defined below.

Definition 1 (*Improved policy*)**.** *For any two policy profiles π and $\bar{\pi}$, we write $I(\bar{\pi}, \pi)$ for the 'improvement' of $\bar{\pi}$ over π, defined as*

$$I(\bar{\pi}, \pi) \stackrel{\text{def}}{=} \sum_{i\in\{1,2\}} V_i^{(\bar{\pi}_i, \pi_{-i})} - V_i^{(\pi_i,\pi_{-i})} = \sum_{i\in\{1,2\}} V_i^{(\bar{\pi}_i,\pi_{-i})}.$$

We say that a policy $\bar{\pi}$ improves over π if $I(\bar{\pi}, \pi) \geq 0$.

2.1. Mirror Ascent against an Improved Opponent

We now introduce Mirror Ascent against an Improved Opponent (MAIO). Consider a strongly convex and continuously-differentiable function $\varphi : \Omega \to \mathbb{R}$, called the regularizer, where the domain $\Omega \subset \mathbb{R}^{|A|}$ contains the simplex $\Delta(A)$, and write D_φ the associated Bregman divergence: for $y, y' \in \Omega$,

$$D_\varphi(y, y') \stackrel{\text{def}}{=} \varphi(y) - \varphi(y') - \nabla\varphi(y') \cdot (y - y').$$

The **MAIO algorithm** defines a sequence of policies $(\pi_{i,t})_{t\geq 0}$ as follows: for all $i \in \{1,2\}$, $\pi_{i,0}$ is the uniform policy, and for all $t \geq 0$,

$$\pi_{i,t+1} \in \operatorname*{arg\,max}_{\pi_i\in\Delta(A_i)} \Big[\eta_t\, \pi_i \cdot Q_i^{\bar{\pi}_{-i,t}} - D_\varphi(\pi_i, \pi_{i,t})\Big], \qquad (1)$$

where $\eta_t > 0$ is a learning rate. For each player i, this is a mirror-ascent step (Nemirovski and Yudin, 1983; Bubeck, 2015; Lattimore and Szepesvári, 2020) on the value $\pi_i \mapsto \pi_i \cdot Q_i^{\bar{\pi}_{-i,t}} = V_i^{(\pi_i,\bar{\pi}_{-i,t})}$ of the policy π_i playing against the improved opponent $\bar{\pi}_{-i,t}$ regularized by $D_\varphi(\pi_i, \pi_{i,t})$, which penalize policies away from the previous policy $\pi_{i,t}$. This definition corresponds to the so-called proximal or trust region view of mirror-descent (MD). Alternatively, an equivalent definition is given in terms of the mirror map $\nabla\varphi : \Omega \to \mathbb{R}^{|A_i|}$ (see e.g., (Bubeck, 2015)):

$$\pi_{i,t+1} = \operatorname*{arg\,min}_{\pi_i\in\Delta(A_i)} D_\varphi(\pi_i, y_{i,t+1}),$$

where $y_{i,t+1}$ is the (unique) point of $\mathbb{R}^{|A_i|}$ such that

$$\nabla\varphi(y_{i,t+1}) = \nabla\varphi(\pi_{i,t}) + \eta_t Q_i^{\bar{\pi}_{-i,t}}.$$

Specifically, a gradient descent step is performed in the mirror space (by application of the mirror map $\nabla\varphi$). Under some assumptions (see e.g. Lattimore and Szepesvári (2020)), MD is equivalent to *Follow the Regularized Leader (FTRL)*. Intuitively, here FTRL would accumulate the Q-values of the improved opponent and derive the policy as a regularized projection step:

$$\pi_{i,t+1} \in \operatorname*{arg\,max}_{\pi_i\in\Delta(A_i)} \Big[\pi_i \cdot \sum_{s=0}^{t} \eta_s Q_i^{\bar{\pi}_{-i,s}} - \varphi(\pi_i)\Big].$$

We now consider two natural choices of regularizers, the entropy regularizer (for which MD is equivalent to FTRL) and the ℓ_2-regularizer (for which it is not).

2.2. Entropy regularization

For the negative entropy regularization $\varphi(\pi) \stackrel{\text{def}}{=} \sum_a \pi(a) \log \pi(a)$ the domain Ω is the interior of $\Delta(A)$ and the Bregman divergence is the KL divergence: $D_\varphi(\pi, \pi') = KL(\pi, \pi') = \sum_a \pi(a) \log \frac{\pi(a)}{\pi'(a)}$. Thus MAIO produces the sequence of policies $\pi_{i,t+1}(a) \propto \pi_{i,t}(a) \exp\big(\eta_t Q_i^{\bar{\pi}_{-i,t}}(a)\big)$. In this case, MD coincides with FTRL, and the policy is the softmax of the accumulated values: $\pi_{i,t+1}(a) \propto \exp\big(\sum_{s=0}^t \eta_s Q_i^{\bar{\pi}_{-i,s}}(a)\big)$.

2.3. ℓ_2-regularization

For the ℓ_2-regularization $\varphi(\pi) \stackrel{\text{def}}{=} \frac{1}{2}\|\pi\|_2^2 = \frac{1}{2}\sum_a \pi(a)^2$, and the domain $\Omega = \mathbb{R}^{|A|}$, the mirror map is the identity ($\nabla\varphi(\pi) = \pi$) and the Bregman divergence is half the square Euclidean norm $D_\varphi(\pi, \pi') = \frac{1}{2}\|\pi - \pi'\|_2^2$. MAIO produces the policies:

$$\begin{aligned}\pi_{i,t+1} &= \operatorname*{arg\,max}_{\pi_i\in\Delta(A_i)} \Big[\eta_t \pi_i \cdot Q_i^{\bar{\pi}_{-i,t}} - \frac{1}{2}\|\pi_i - \pi_{i,t}\|_2^2\Big] \\ &= \operatorname*{arg\,min}_{\pi_i\in\Delta(A_i)} \big\|\pi_i - \big(\pi_{i,t} + \eta_t Q_i^{\bar{\pi}_{-i,t}}\big)\big\|_2^2\end{aligned}$$

which is the projected gradient descent algorithm:

$$\pi_{i,t+1} = P_{\Delta(A_i)}\big(\pi_{i,t} + \eta_t Q_i^{\bar{\pi}_{-i,t}}\big),$$

where $P_{\Delta(A_i)}$ is the ℓ_2-projection onto the simplex $\Delta(A_i)$ (also called sparsemax operator, see e.g., (Martins and Astudillo, 2016), because it induces sparsity). Notice that this

algorithm is different from a FTRL (with ℓ_2 regularization) version of the algorithm, which would be defined as

$$\pi_{i,t+1} \quad = \quad P_{\Delta(A_i)}\big(\sum_{s=0}^{t} \eta_s Q_i^{\bar{\pi}_{-i,s}}\big).$$

The results we present in the next section apply to the MD version; it is an open question to whether similar results could be obtained with the FTRL version.

2.4. Convergence to the set of Nash Equilibria

First, we recall that φ is a strongly convex function with respect to some norm $\|\cdot\|$ and with modulus σ, if for any $y, y' \in \Omega$,

$$\varphi(y) \geq \varphi(y') + \nabla\varphi(y') \cdot (y - y') + \frac{\sigma}{2}\|y - y'\|^2. \quad (2)$$

In the two cases we have considered previously, we have that the ℓ_2-regularizer $\varphi(\pi) = \frac{1}{2}\|\pi\|^2$ is strongly convex w.r.t. ℓ_2-norm with modulus $\sigma = 1$, and the entropy regularizer $\varphi(\pi) = \sum_a \pi(a)\log\pi(a)$ is strongly convex w.r.t. ℓ_1-norm with modulus $\sigma = 1$ (from Pinsker's inequality, see e.g., Csiszar and Korner (1982)).

For a given regularizer φ, we write J_{π^*} the Bregman divergence between any policy π and a Nash equilibrium π^*:

$$J_{\pi^*}(\pi) \stackrel{\text{def}}{=} \sum_{i\in\{1,2\}} D_\varphi(\pi_i^*, \pi_i).$$

The main property of MAIO is that at each iteration this distance to any Nash eq. decreases as a function of how much the policy $\bar{\pi}_t$ improves over the current policy π_t.

Theorem 1. *Let $\pi^* \in \Pi^*$ be any Nash equilibrium. Let φ be a strongly convex function w.r.t. the ℓ_p-norm with modulus σ, and let $q = 1/(1-1/p)$. MAIO builds a sequence of policies (π_t) defined by (1) such that*

$$J_{\pi^*}(t+1) \quad \leq \quad J_{\pi^*}(t) - \eta_t I(\bar{\pi}_t, \pi_t) + c\eta_t^2,$$

where $J_{\pi^*}(t) \stackrel{\text{def}}{=} J_{\pi^*}(\pi_t)$, $c \stackrel{\text{def}}{=} \frac{4}{\sigma}|A|^{2/q}Q_{\max}^2$ and $Q_{\max}$ is the maximum absolute entry of the reward matrix R.

In particular, with the choice $\eta_t = \frac{I(\bar{\pi}_t,\pi_t)}{2c}$, we have

$$J_{\pi^*}(t+1) \quad \leq \quad J_{\pi^*}(t) - \frac{I(\bar{\pi}_t, \pi_t)^2}{4c}.$$

This result says that as long as we can find a policy $\bar{\pi}_t$ which improves over the current policy π_t (in the sense of $I(\bar{\pi}_t, \pi_t) > 0$), then MAIO produces a policy π_{t+1} which is closer to any Nash equilibrium than the previous policy. Since we know that the set of policies which cannot be improved are the set of Nash equilibria (by definition of the improvement I), we deduce that the speed at which MAIO converges to the set of Nash equilibria depends on how much the policies $\bar{\pi}_t$ improve over π_t.

In the next sub-section we consider the best response as improved policy.

2.5. MAIO with the best response

The policy $\bar{\pi}$ which improves the most over π (in the sense of maximizing $\bar{\pi} \mapsto I(\bar{\pi}, \pi)$) is the best response, i.e.

$$b(\pi) \stackrel{\text{def}}{\in} \arg\max_{\bar{\pi}} I(\bar{\pi}, \pi) = \arg\max_{(\bar{\pi}_1,\bar{\pi}_2)} \big(V^{(\bar{\pi}_1,\pi_2)} - V^{(\pi_1,\bar{\pi}_2)}\big).$$

We now show that the improvement of the best response over any policy π is lower-bounded by the ℓ_2-distance between π and the set of Nash equilibria.

Define $I^*(\pi) \stackrel{\text{def}}{=} I(b(\pi), \pi) = \max_{\bar{\pi}} I(\bar{\pi}, \pi)$ to be the improvement of the best response over policy π, also called **exploitability**, see (Ponsen et al., 2011).

Lemma 1. *There exists a constant $\kappa > 0$ (which depends on the matrix R only) such that for any policy π we have*

$$I^*(\pi) \quad \geq \quad \kappa \min_{\pi^*\in\Pi^*} \|\pi - \pi^*\|_2,$$

where the norm between policy profiles is $\|\pi - \pi'\|_2 \stackrel{\text{def}}{=} \big(\sum_{i\in\{1,2\}} \|\pi_i - \pi_i'\|_2^2\big)^{1/2}$

This result, combined with Theorem 1 with the ℓ_2 regularizer, implies that MAIO using the best response as improved opponent (MAIO-BR) converges to the set of Nash equilibria with an exponential rate:

Theorem 2. *Consider the MAIO-BR algorithm (MAIO with best response as improved opponent) with ℓ_2-regularizer. Choose the learning rate $\eta_t = \frac{I(\bar{\pi}_t,\pi_t)}{2c}$. Then*

$$\min_{\pi^*\in\Pi^*} \|\pi^* - \pi_t\|_2 \quad \leq \quad e^{-\beta t} \min_{\pi^*\in\Pi^*} \|\pi^* - \pi_0\|_2,$$

with $\beta \stackrel{\text{def}}{=} \kappa^2/(16|A|Q_{\max}^2)$.

Proof. From Theorem 1 (with $p = q = 2$ and choosing φ to be the ℓ_2 regularizer) we have, for any $\pi^* \in \Pi^*$, $J_{\pi^*}(t) = \frac{1}{2}\|\pi^* - \pi_t\|_2^2$, thus, with $c = 4|A|Q_{\max}^2$,

$$\begin{aligned}\|\pi^* - \pi_{t+1}\|_2^2 \quad &\leq \quad \|\pi^* - \pi_t\|_2^2 - \frac{I^*(\pi_t)^2}{2c} \\ &\leq \quad \|\pi^* - \pi_t\|_2^2 - \frac{\kappa^2}{2c} \min_{\pi^*\in\Pi^*} \|\pi^* - \pi_t\|_2^2,\end{aligned}$$

where the last inequality comes from Lemma 1. Taking the minimum over Π^*,

$$\min_{\pi^*\in\Pi^*} \|\pi^* - \pi_{t+1}\|_2^2 \quad \leq \quad \min_{\pi^*\in\Pi^*} \|\pi^* - \pi_t\|_2^2 \Big(1 - \frac{\kappa^2}{2c}\Big).$$

Thus $\min_{\pi^*\in\Pi^*} \|\pi^* - \pi_t\|^2$ decreases exponentially fast and the result holds with $\beta = \kappa^2/(4c) = \kappa^2/(16|A|Q_{\max}^2)$. □

2.6. Interpretation of κ

Lemma 1 yields the existence of a constant $\kappa > 0$ that controls the exponential rate of convergence of the MAIO-BR to the set of Nash equilibria. Intuitively, κ measures the flatness of exploitability function $\pi \mapsto I^*(\pi)$ near the set of Nash equilibria. More precisely κ is a lower bound on directional derivatives of $I^*(\pi)$ for $\pi \notin \Pi^*$. This quantity also appears in the analysis of the first-order smoothing method due to Gilpin et al. (2008; 2012), with detailed analysis of the quantity itself appearing in Mordukhovich et al. (2010). We also provide an interpretable lower bound on κ in Appendix D.

Example 1. *To get some intuition about κ, let us consider the simple game defined by the reward matrix $R = \begin{pmatrix} 0 & 1-\varepsilon \\ 2 & 1 \end{pmatrix}$. The Nash eq. is $\pi_1^* = (0,1)$ and $\pi_2^* = (0,1)$ and the game has a minimax value of 1. We can prove that the derivative of $\pi_1 \mapsto I^*((\pi_1, \pi_2))$ around π^* is lower bounded by ε and that $I^*(\pi) \geq \kappa \|\pi - \pi^*\|_2$ for $\kappa = \varepsilon/\sqrt{2}$ (see Appendix E). And indeed, numerical results show that MAIO-BR's exponential convergence to the Nash eq. depends on the value of ε (see Section 4).*

Remark 1. *We achieve exponential convergence rate using the ℓ_2 regularization. An interesting question is whether an exponential convergence is achieved in the case of entropy regularization as well. We conjecture that this is true if and only if (at least) one Nash eq. is an interior point (strictly stochastic policy). See some arguments for this conjecture in Appendix O and the experiments in Section 4.*

Remark 2. *Our results concern the distance to the Nash eq. in policy space, rather than in value space, which explains the dependence of the bounds on κ, which encodes the flatness of the exploitability function close to the set of Nash equilibria. In general, bounds on policy distance can be straightforwardly translated to and from bounds on value approximation via multiplication by game-dependent constants, such as κ and the maximum spread of rewards available in the game (see Lemma 2 for the IIG case).*

We now present the extension of MAIO to IIGs.

3. Sequential Imperfect Information Games

3.1. Notations

In the setting of imperfect information games (IIGs) in sequential form, we assume the players $\{1, 2\}$ play sequentially. The case of simultaneous actions could be handled via non-observability of the opponent's actions. Let $H = \cup_{i \in \{1,2\}} H_i$ be the set of possible histories, with H_i being the histories from which player $i \in \{1, 2\}$ may play. Similarly let $X = \cup_{i \in \{1,2\}} X_i$ be the set of observations (also called states or information nodes). We assume a deterministic observation process and use set notation to represent an observation $x(h)$ that corresponds to a set of possible histories $h \in x$. For any $h \in H$, we denote by $i(h) \in \{1, 2\}$ the player whose turn it is to play in h, i.e. $h \in H_{i(h)}$.

We write $p(h'|h, a)$ for the (sub-)probability of transitioning from $h \in H_i$ to h' when player $i = i(h)$ selects action $a \in A_i$ in h. The initial history h_0 is drawn from some initial distribution ρ_0 and we assume a terminal state $\emptyset$ from which there is no reward. At each transition, the probability of reaching this terminal state is $p(\emptyset|h, a) = 1 - \sum_{h'} p(h'|h, a)$. This setting covers stochastic shortest path (for which it is assumed that for any policy the expected time to reach $\emptyset$ is finite), finite-time horizon (probability to reach $\emptyset$ is 1 when the time horizon is reached, otherwise 0), and discounted infinite horizon problems (probability to reach $\emptyset$ is $1 - \gamma$ at every transition, where $\gamma < 1$ is the discount factor). We assume the underlying process at the history level H is Markovian with a tree structure (i.e., there exists a unique path from h_0 to any history $h \in H$) and that the history and action spaces are finite.

Actions are drawn from the player's policy $\pi_i : X_i \to \Delta(A_i)$ and are a function of the observations. We write $\pi = \{\pi_i\}_{i \in \{1,2\}} = (\pi_i, \pi_{-i})$ the policy whose restriction to X_i is π_i.

Finally, the reward function for each player i is denoted by $r_i(h, a)$ and is assumed to be a deterministic function of the history and action. The game is zero-sum thus $r_i = -r_{-i}$.

3.2. Reach probabilities and value function

History reach probabilities: We define the probability of reaching a history h under a policy profile π as

$$\mu^\pi(h) \stackrel{\text{def}}{=} \mathbb{E}_{h_0 \sim \rho_0}\Big[\sum_{k \geq 0} \mathbb{I}\{h_k = h\}\Big],$$

where $(h_k)_{k \geq 0}$ is the Markov chain on H induced by the policy π. These reach probabilities satisfy the balance equation:

$$\mu^\pi(h') = \rho_0(h') + \sum_{h \in H} \mu^\pi(h) \sum_a \pi(a|h) p(h'|h, a), \quad (3)$$

where $\pi(a|h) \stackrel{\text{def}}{=} \pi_{i(h)}(a|x(h))$.

Observation reach probabilities: We define the probability of an observation x as $\mu^\pi(x) \stackrel{\text{def}}{=} \sum_{h \in x} \mu^\pi(h)$.

History-based value functions: We define the history-based Q-function, for $h \in H_i$, $a \in A_i$,

$$Q_i^\pi(h, a) = \mathbb{E}\Big[\sum_{k \geq 0} r_i(h_k, a_k) | h_0 = h, a_0 = a\Big], \quad (4)$$

and the state value function:

$$V_i^\pi(h) = \mathbb{E}\Big[\sum_{k\geq 0} r_i(h_k, a_k) | h_0 = h\Big] = \sum_{a\in A_i} \pi(a|h) Q_i^\pi(h,a).$$

We define the initial value function V_i^π as the value of the game for player i:

$$V_i^\pi \stackrel{\text{def}}{=} \mathbb{E}_{h_0\sim\rho_0}[V_i^\pi(h_0)]. \tag{5}$$

Using the reach probabilities, we have

$$V_i^\pi = \sum_{h\in H} \mu^\pi(h) \sum_a \pi(a|h) r_i(h,a). \tag{6}$$

Value function on observations: For any state x such that $\mu^\pi(x) > 0$, we define its Q-value as the convex combination of the Q-value of the corresponding histories $h \in x$ weighted by their conditional probability $\mu^\pi(h|x) \stackrel{\text{def}}{=} \frac{\mu^\pi(h)}{\mu^\pi(x)}$:

$$Q_i^\pi(x,a) \stackrel{\text{def}}{=} \sum_{h\in x} \mu^\pi(h|x) Q_i^\pi(h,a). \tag{7}$$

Thus $Q_i^\pi(x,a)$ depends on the policy π both in terms of the future reward collected when following π from x on, but also in terms of the probabilities $\mu^\pi(h)$ of reaching specific histories $h \in x$ when following π.

3.3. Perfect recall

The reach probability of any history $\mu^\pi(h)$ is the product along the path $(h_0, a_0, h_1, a_1, \ldots, h_{n-1}, a_{n-1}, h_n = h)$, for some $n \geq 0$ (n is the depth of the history h), of the action probabilities $\pi(a_k|h_k)$ and the transition probabilities $p(h_{k+1}|h_k, a_k)$, for $k \leq n$. Factorizing the probabilities per player, we write

$$\mu^\pi(h) = \mu_0(h) \prod_{i\in\{1,2\}} \mu_i^\pi(h),$$

where $i \in \{1,2\}$ corresponds to player's i policy: $\mu_i^\pi(h) \stackrel{\text{def}}{=} \prod_{k=0\ldots n-1 : i(h_k)=i} \pi(a_k|h_k)$, and μ_0 corresponds to the transition probabilities: $\mu_0(h) \stackrel{\text{def}}{=} \rho_0(h_0) \prod_{k=0\ldots n-1} p(h_{k+1}|h_k, a_k)$.

We now make the so-called *perfect recall* assumption that for each player i, any information node $x \in X_i$ contains all information about previous information nodes for player i as well as its past actions:

Assumption 1 (Perfect recall)**.** *For each player $i \in \{1,2\}$, all $x \in X_i$, all $h, h' \in x$, any policy π, we assume that $\mu_i^\pi(h) = \mu_i^\pi(h')$.*

Under this assumption we can define $\mu_i^\pi(x) \stackrel{\text{def}}{=} \mu_i^\pi(h)$, for $x = x(h)$. As a consequence, the reach probability $\mu^\pi(x) = \sum_{h\in x} \mu^\pi(h)$ of any observation $x \in X_i$ can be factorized as the product of $\mu_i^\pi(x)$ (Player i's contribution to reach x) and $\mu_{\neq i}^\pi(x)$ (the opponent's and chance's contributions to reach x):

$$\mu^\pi(x) = \sum_{h\in x} \mu_0(h) \mu_i^\pi(h) \mu_{-i}^\pi(h) = \mu_i^\pi(x) \mu_{\neq i}^\pi(x),$$

where $\mu_{\neq i}^\pi(x) \stackrel{\text{def}}{=} \sum_{h\in x} \mu_{\neq i}^\pi(h)$ and $\mu_{\neq i}^\pi(h) \stackrel{\text{def}}{=} \mu_0(h)\mu_{-i}^\pi(h)$.

We deduce that for any two policy profiles π and π', $x \in X_i$,

$$\mu^{(\pi_i, \pi'_{-i})}(x) = \mu_i^\pi(x) \mu_{\neq i}^{\pi'}(x). \tag{8}$$

The MDP $\mathcal{M}_i^{\pi_{-i}}$: In general, the observation process $(x_t = x(h_k))_{k\geq 0}$ is a POMDP. However, under the perfect recall assumption, if we fix the policy π_i of the opponent, then the observation process $(x_k)_{k\geq 0 : i(h_k)=i}$ (at successive times k when it is Player i's turn to play) forms an MDP, which we write as $\mathcal{M}_i^{\pi_{-i}}$. In particular, the probability to transit from $x \in X_i$ to another $x' \in X_i$ does not depend on the player's own policy. See Proposition 1 and Section I in the Appendix for the precise definition and properties of this MDP.

3.4. MAIO for Imperfect Information Games

MAIO requires being able to compute an improved policy $\bar{\pi}_t$ over π_t at each iteration. The improvement $I(\bar{\pi}_t, \pi_t)$ is defined exactly as in Definition 1 where the value functions V_i^π are considered from the initial state (5).

Algorithm [MAIO for IIG]: For each player $i \in \{1,2\}$, we start with a uniform policy $\pi_{i,0}(x)$ from all $x \in X_i$. At every iteration $t \geq 0$, we compute an improved policy $\bar{\pi}_t(x)$ over π_t (several possible choices are described later). For each player i, we evaluate the Q-values $Q_i^{(\pi_{i,t}, \bar{\pi}_{-i,t})}(h,a)$ and reach probabilities $\mu_{\neq i}^{\bar{\pi}_t}(h)$ and we define a new policy π_{t+1}, for each $x \in X_i$, as

$$\pi_{i,t+1}(x) \in \underset{\pi_i\in\Delta(A_i)}{\arg\max} \Big[-D_\varphi(\pi_i, \pi_{i,t}(x)) + \eta_t \sum_{a\in A_i} \pi_i(a) \sum_{h\in x} \mu_{\neq i}^{\bar{\pi}_t}(h) Q_i^{(\pi_{i,t}, \bar{\pi}_{-i,t})}(h,a) \Big]. \tag{9}$$

Notice that if $\mu_i^{(\pi_{i,t}, \bar{\pi}_{-i,t})}(x) \neq 0$ then (see (15) for a proof),

$$\sum_{h\in x} \mu_{\neq i}^{\bar{\pi}_t}(h) Q_i^{(\pi_{i,t}, \bar{\pi}_{-i,t})}(h,a) = \mu_{\neq i}^{\bar{\pi}_t}(x) Q_i^{(\pi_{i,t}, \bar{\pi}_{-i,t})}(x,a).$$

We notice that this MAIO algorithm for IIGs makes use of the counterfactual reach probabilities $\mu_{\neq i}^{\bar{\pi}_t}(h)$ introduced in counterfactual regret minimization algorithms (Zinkevich et al., 2008).

Now we analyze the theoretical properties of this algorithm.

3.5. Theoretical analysis of MAIO-IIG

Letting $\pi^* \in \Pi^*$ be any Nash eq. of the game, we introduce the energy function of the IIG:

$$J_{\pi^*}(\pi) \quad \stackrel{\text{def}}{=} \quad \sum_{i\in\{1,2\}} \sum_{x\in X_i} \mu_i^{\pi^*}(x) D_\varphi(\pi_i^*(x), \pi_i(x)),$$

and we write $J_{\pi^*}(t) = J_{\pi^*}(\pi_t)$. Our main result is the following:

Theorem 3. *The MAIO algorithm for IIG produces a sequence of policies such that for any Nash eq. $\pi^* \in \Pi^*$,*

$$J_{\pi^*}(t+1) \quad \leq \quad J_{\pi^*}(t) - \eta_t I(\bar{\pi}_t, \pi_t) + c\eta_t^2,$$

where $c \stackrel{\text{def}}{=} \frac{4}{\sigma}|A|^{2/q} Q_{\max}^2 L_{\max}$, $Q_{\max} = \max_\pi \max_{h\in H, a\in A} |Q^\pi(h,a)|$, and $L_{\max} = \max_\pi \sum_x \mu^\pi(x)$. Thus, with $\eta_t = \frac{I(\bar{\pi}_t,\pi_t)}{2c}$ we have

$$J_{\pi^*}(t+1) \quad \leq \quad J_{\pi^*}(t) - \frac{I(\bar{\pi}_t,\pi_t)^2}{4c}.$$

Remark 3. *The coefficient $L_{\max}$ is a bound on the effective time horizon (in the case of finite horizon, $L_{\max}$ is a lower bound on the time horizon, in the discounted setting $L_{\max} = 1/(1-\gamma)$ and in the case of stochastic shortest path problems it is the largest expected time before reaching the terminal state).*

This result is similar to Theorem 1 in the sense that it states that the current policy gets closer to the Nash eq. as long as $\bar{\pi}_t$ improves over π_t. The distance to the Nash eq. is measured in terms of $J_{\pi^*}(\pi)$ which is a distance in policy space. More precisely, $J_{\pi^*}(\pi)$ measures the Bregman divergence between the policy $\pi_t(x)$ and $\pi^*(x)$ weighted by the player i's own probability $\mu_i^{\pi^*}(x)$ to reach $x \in X_i$ when following a Nash eq. policy.

Now in the IIG setting, there are several ways to compute an improved policy $\bar{\pi}_t$ which will be discussed later. First we consider as improved policy, the best response, which provides the largest improvement.

3.6. MAIO-BR for IIG

Now, let us consider as improved policy the best response, i.e., $b_{i,t} \in \arg\max_\pi I(\pi, \pi_t)$. First we show that the exploitability $I^*(\pi)$ of any policy π is upper bounded by its ℓ_2-energy distance J to Π^*. Thus minimizing the J-distance to the set of Nash eq. implies minimizing exploitability as well.

Lemma 2. *For any policy π, we have*

$$I^*(\pi)^2 \quad \leq \quad L_{\max}^2 |A| Q_{\max}^2 \min_{\pi^*\in\Pi^*} J_{\pi^*}(\pi).$$

Now we state a convergence result to the set of Nash eq.

Theorem 4 (Convergence of MAIO-BR)**.** *The sequence of policies produced by MAIO-BR algorithm with $\eta_t = I^*(\pi_t)/(2c)$ converges to the set of Nash equilibria, in the sense that $\lim_{t\to\infty} \min_{\pi^*\in\Pi^*} J_{\pi^*}(\pi_t) = 0$. Notice that from Lemma 2 we also deduce the result in exploitability: $\lim_{t\to\infty} I^*(\pi_t) = 0$.*

In the normal form games we could deduce an exponential convergence speed to the set of Nash eq. thanks to Lemma 1. Unfortunately, in the case of IIGs, we do not have a similar result. Indeed we have the following counter-example:

Lemma 3. *There exists a two-player zero-sum imperfect information game such that there exists no $\kappa > 0$ such that for all π, $I^*(\pi) \geq \kappa \min_{\pi^*\in\Pi^*} \sqrt{J_{\pi^*}(\pi)}$, where J_{π^*} is the energy distance (i.e. φ is the ℓ_2 norm).*

The reason why the situation in IIGs is different from that in normal form games is that the mapping $\pi_i \mapsto V_i^{(\pi_i,\pi_{-i})}$ is not globally linear in π_i.

However, under the perfect recall assumption, the value function is linear w.r.t. the individual reach probability of each player (the so-called sequence form, see e.g. (Von Stengel, 1996)). Thus by defining the ℓ_2 distance in reach probabilities:

$$d(\pi, \pi') \quad \stackrel{\text{def}}{=} \quad \sum_{i\in\{1,2\}} \sum_{x\in X_i, a\in A} \left[\mu_i^\pi(x,a) - \mu_i^{\pi'}(x,a)\right]^2,$$

where we write $\mu_i^\pi(x,a) \stackrel{\text{def}}{=} \mu_i^\pi(x)\pi_i(a|x)$, we can deduce the following result:

Lemma 4. *There exists a constant $\kappa > 0$ (which depends on the game), such that for any policy π we have*

$$I^*(\pi) \geq \kappa \min_{\pi^*\in\Pi^*} \sqrt{d(\pi^*, \pi)}$$

We can also show that the distance $J_{\pi^*}(\pi) = O(d(\pi^*, \pi))$:

Lemma 5. *For any π^* there exists two constant $\delta, c > 0$ such for any π such that $d(\pi^*, \pi) \leq \delta$, we have*

$$J_{\pi^*}(\pi) \leq c\, d(\pi^*, \pi).$$

Notice that Lemmas 5 and 4 do not contradict Lemma 3 because the constants δ and c in Lemmas 5 depend on the specific choice of the policy $\pi^* \in \Pi^*$.

Now, under some assumption of the set of Nash eq., we can combine Lemmas 5 and 4 together with Theorem 3 to deduce an exponential rate of convergence.

Theorem 5. *Consider MAIO-BR with a ℓ_2-regularizer, and a learning rate $\eta_t = \frac{I^*(\pi_t)}{2c}$. Define*

$$\varepsilon = \inf_{\pi^*\in\Pi^*, i\in\{1,2\}, x\in X_i, a\in A: \mu_i^{\pi^*}(x,a)>0} \mu_i^{\pi^*}(x,a).$$

If $\varepsilon > 0$ then MAIO-BR converges to the set of Nash equilibria at an exponential rate.

Notice that a sufficient condition for $\varepsilon > 0$ (thus in order that MAIO-BR enjoys an exponential rate) is that the Nash eq. is unique.

3.7. Improved policies

MAIO for IIG requires computing an improved policy $\bar{\pi}$ over the current one π. In the case of IIGs there are several possible choices for computing such improved policies with different trade-off between computational complexity versus amount of improvement, thus speed of convergence to the Nash eq. Here are a few examples. First we introduce the notion of local improvement and derive a sufficient condition for a policy $\bar{\pi}$ to improve over π.

Define the **local improvement**: for any $x \in X_i$,

$$I_i(\bar{\pi}, \pi)(x) \stackrel{\text{def}}{=} \sum_{a \in A_i} (\bar{\pi}_i(a|x) - \pi_i(a|x)) Q_i^{(\pi_i, \pi_{-i})}(x, a).$$

Lemma 6. *Given two policy profiles π and $\bar{\pi}$. If, for any Player i, any $x \in X_i$, the local improvement $I_i(\bar{\pi}, \pi)(x) \geq 0$, then $\bar{\pi}$ improves over π, i.e., $I(\bar{\pi}, \pi) \geq 0$. In addition, if $I_i(\bar{\pi}, \pi)(x) > 0$ for some $x \in X_i$ such that $\mu^{(\bar{\pi}_i, \pi_{-i})}(x) > 0$, then $I(\bar{\pi}, \pi) > 0$.*

Proof. Applying Lemma 9 (in the Appendix) to the policies $\pi_i, \bar{\pi}_i$, and π_{-i}, the improvement is

$$I(\bar{\pi}, \pi) \quad = \quad \sum_i \sum_{x \in X_i} \mu^{(\bar{\pi}_i, \pi_{-i})}(x) I_i(\bar{\pi}, \pi)(x),$$

from which we deduce our claim. □

This result tells us that in order to find an improved policy $\bar{\pi}$ it is sufficient that from each state $x \in X_i$, the expected $Q_i^{\pi}(x, \cdot)$-values under policy $\bar{\pi}_i(\cdot|a)$ are larger than under the current policy $\pi_i(\cdot|a)$. Here are a few examples of improved policies.

Best response: for each i, $b_i = \arg\max_{\pi_i'} V^{(\pi_i', \pi_{-i})}$. This is the policy which improves the most. In this case $I(b, \pi)$ represents the exploitability of the current policy, and we have seen in Theorem 5 that an exponential rate of convergence can be achieved. However computing the best response at each iteration is computationally expensive as it requires solving an optimal control problem, so we may prefer cheaper alternatives.

Greedy policy: The greedy policy is easy to deduce once the Q-values of the current policy have been computed: $g_i(x) \stackrel{\text{def}}{\in} \arg\max_a Q_i^{\pi}(x, a)$. From Lemma 6 this policy provides an improvement over π, thus $I(g, \pi) \geq 0$. However it is possible that $I(g, \pi) = 0$ while π is not a Nash eq. yet, see Appendix P for an illustration of this situation and several solutions to circumvent this problem. Computing a greedy policy has a smaller computational complexity than computing the best response since it requires evaluating a fixed policy instead of finding the optimal one.

Optimistic mirror descent and extra-gradient method: (see e.g., (Mertikopoulos et al., 2019)) one could follow a step of mirror descent against the current opponent which, in the IIG setting here, would correspond to defining the improved policy as

$$\bar{\pi}_{i,t}(x) \in \arg\max_{\pi_i \in \Delta(A_i)} \Big[\rho_t \mu_{\neq i}^{\pi}(x) \; \pi_i \cdot Q_i^{\pi}(x) - D_{\varphi}(\pi_i, \pi_{i,t}(x)) \Big],$$

for some step $\rho_t > 0$. It is possible to prove that this policy $\bar{\pi}_t$ improves locally over the current policy π_t: $I(\bar{\pi}_t, \pi_t)(x) = D_{\varphi}(\bar{\pi}_t(x), \pi_t(x))$ thus improves globally as well, from Lemma 6.

Mixture policy: Any mixture between an improved policy $\bar{\pi}$ and the current policy π improves over the current policy. For example one could use the mixture $\bar{\pi}^{\alpha} \stackrel{\text{def}}{=} (1 - \alpha)\pi + \alpha\bar{\pi}$ between the current and improved policies, defined for every $x \in X_i$ as

$$\bar{\pi}_i^{\alpha}(a|x) \propto (1 - \alpha)\mu_i^{\pi}(x)\pi_i(a|x) + \alpha\mu_i^{\bar{\pi}}(x)\bar{\pi}_i(a|x), \quad (10)$$

(see e.g. Heinrich et al. (2015) Lemma 6, or Zinkevich et al. (2008) Eq. (4)). The value function of this mixture is the convex combinations of the value functions: $V_i^{(\bar{\pi}_i^{\alpha}, \pi_{-i})} = (1 - \alpha)V_i^{(\pi_i, \pi_{-i})} + \alpha V_i^{(\bar{\pi}_i, \pi_{-i})}$. Thus the improvement of this mixture is $I(\bar{\pi}^{\alpha}, \pi) = \alpha I(\bar{\pi}, \pi)$. A possible benefit of using this mixture for small α is that this policy is close to the current policy, so we can think of using off-policy techniques in sampling-based policy evaluation algorithms, while guaranteeing convergence to the Nash eq.

MCTS improved policy: An improved policy could be obtained by Monte Carlo Tree Search (or any other planning algorithm). This would return an improved policy whose improvement depends on the depth of the search, from the greedy policy (corresponding to 1-step look-ahead search) to the full best response (full tree search). Thus the MAIO setting allows one to use MTCS for computing Nash eq. in IIGs. The trade-off is computational complexity (as a function of the depth of the search) versus the amount of improvement (thus how fast the algorithm converges to the Nash eq.) of the policy returned by the search.

4. Numerical experiments on matrix games

Here we evaluate MAIO-BR on 2 matrix games with both ℓ_2 and entropy regularization. In the Appendix, Section P we report experiments of MAIO for IIG and compare to other approaches (CFR, CFR-BR, CFR+).

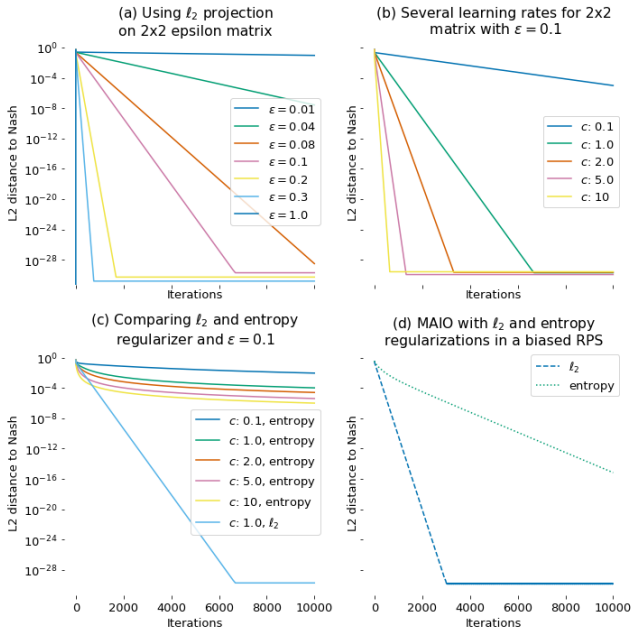

Figure 1. We report ℓ_2 distance to the Nash eq. (in log-scale) for MAIO-BR on the ε-matrix game (Fig. a,b,c) and the biased rock-paper-scissors game (Fig. d). MAIO-BR with ℓ_2 regularization shows an exponential convergence whose rate depend on ε (Fig. a) and the constant c (Fig. b) used in the learning rate. On the contrary, MAIO-BR with softmax does not enjoy an exponential rate (Fig. c) since the Nash eq. is deterministic. However in a games where the Nash eq. is interior, both ℓ_2 and soft-max show exponential convergence (Fig. d). Non-zero value in the plots is explained by numerical precision (we use the `numpy` package with double precision).

The first game is defined by the matrix payoff: $R = \left(\begin{smallmatrix} 0 & 1-\varepsilon \\ 2 & 1 \end{smallmatrix}\right)$ parameterized by some $\varepsilon > 0$. See the discussion in subsection 2.6 (and Appendix E). The ℓ_2 distance to the Nash eq. is reported in Figure 1. We observe the exponential convergence with a rate that depends on ε (Fig. 1(a)) and the constant c (Fig. 1(b)) used in the learning rate (i.e., we chose $\eta_t = c \cdot I(\bar{\pi}_t, \pi_t)$). This is exactly what is predicted by the theory since the value of κ in Lemma 1 is $\varepsilon/\sqrt{2}$ here.

Fig. 1(c) corroborates our conjecture mentioned in subsection 2.6 (see Appendix O) that MAIO-BR with entropy regularization does not enjoy exponential convergence (for any c) when the Nash eq. is not an interior point (here it is a corner of the simplex: $\pi_1^* = (0, 1)$ and $\pi_2^* = (0, 1)$). On the contrary, Fig. 1(d) shows that MAIO-BR enjoys exponential convergence both with ℓ_2 and entropy regularizers (although ℓ_2 seems faster) on the (biased) rock-paper-scissors game, defined by $R = \left(\begin{smallmatrix} 0 & -1 & 0.1 \\ 1 & 0 & -0.1 \\ -0.1 & 0.1 & 0 \end{smallmatrix}\right)$, for which the Nash eq. is interior.

5. Conclusion

We introduced a new class of algorithms for computing a Nash equilibrium in zero-sum normal form games and sequential IIGs and provided an analysis of the speed of convergence in terms of the notion of improvement. We show a new tradeoff between computational complexity of computing improved policies and speed of convergence to the set of Nash eq. Under some condition (including when the Nash eq. is unique) exponential convergence is achieved when we use the best response as improved policy. Maybe the main contribution of MAIO is that it offers a principled approach to use *any* reinforcement learning policy improvement technique (one-step greedy policy, MCTS-improved policy, or even a policy improved by policy gradient) to generate a sequence of policies with convergence guarantee to the set of Nash equilibria.

References

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357.

Chen, Y. and Ye, X. (2011). Projection onto a simplex. *arXiv preprint arXiv:1101.6081*.

Csiszar, I. and Korner, J. (1982). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc.

Daskalakis, C., Deckelbaum, A., and Kim, A. (2011). Near-optimal no-regret algorithms for zero-sum games. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

Daskalakis, C. and Panageas, I. (2018). Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv*.

Gidel, G., Jebara, T., and Lacoste-Julien, S. (2016). Frank-wolfe algorithms for saddle point problems. In *Artificial Intelligence and Statistics (AISTATS)*.

Gilpin, A., Hoda, S., Pena, J., and Sandholm, T. (2007). Gradient-based algorithms for finding Nash equilibria in extensive form games. In *International Workshop on Web and Internet Economics*.

Gilpin, A., Peña, J., and Sandholm, T. (2012). First-order algorithm with $O(\ln(1/\varepsilon))$ convergence for ε-equilibrium in two-person zero-sum games. *Mathematical programming*, 133(1-2):279–298.

Gilpin, A., Peña, J., and Sandholm, T. W. (2008). First-order algorithm with $O(ln(1/\varepsilon))$ convergence for equilibrium in two-person zero-sum games. In *AAAI Conference on Artificial Intelligence*.

Heinrich, J., Lanctot, M., and Silver, D. (2015). Fictitious self-play in extensive-form games. In *International Conference on Machine Learning (ICML)*.

Heinrich, J. and Silver, D. (2016). Deep reinforcement learning from self-play in imperfect-information games. *arXiv*.

Hoda, S., Gilpin, A., Pena, J., and Sandholm, T. (2010). Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512.

Johanson, M., Bard, N., Burch, N., and Bowling, M. (2012). Finding optimal abstract strategies in extensive form games. In *AAAI Conference on Artificial Intelligence*.

Kangarshahi, E. A., Hsieh, Y.-P., Sahin, M. F., and Cevher, V. (2018). Let's be honest: An optimal no-regret framework for zero-sum games. In *International Conference on Machine Learning (ICML)*.

Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. In *ACM Symposium on Theory of Computing (STOC)*.

Khachiyan, L. (1980). Polynomial algorithms in linear programming. *USSR Computational Mathematics and Mathematical Physics*, 20(1):53 – 72.

Koller, D., Megiddo, N., and von Stengel, B. (1994). Efficient solutions of extensive two-person games. In *ACM Symposium on the Theory of Computing (STOC)*.

Koller, D. and Pfeffer, A. (1997). Representations and solutions for game-theoretic problems. *Artificial intelligence*, 94(1-2):167–215.

Korpelevich, G. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

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

Kuhn, H. W. (1950). A simplified two-person poker. *Contributions to the Theory of Games*, 1:97–103.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., Vylder, B. D., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., and Ryan-Davis, J. (2019). OpenSpiel: A framework for reinforcement learning in games. *arXiv*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lockhart, E., Lanctot, M., Pérolat, J., Lespiau, J.-B., Morrill, D., Timbers, F., and Tuyls, K. (2019). Computing approximate equilibria in sequential adversarial games by exploitability descent. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Martins, A. F. T. and Astudillo, R. F. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning (ICML)*.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C., Chandrasekhar, V., and Piliouras, G. (2019). Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations (ICLR)*.

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. (2018). Cycles in adversarial regularized learning. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020). A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *Artificial Intelligence and Statistics (AISTATS)*.

Mordukhovich, B. S., Peña, J. F., and Roshchina, V. (2010). Applying metric regularity to compute a condition measure of a smoothing algorithm for matrix games. *SIAM Journal on Optimization*, 20(6):3490–3511.

Nemirovski, A. and Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics.

Nesterov, Y. (2005). Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249.

Nesterov, Y. E. and Todd, M. J. (1998). Primal-dual interior-point methods for self-scaled cones. *SIAM Journal on Optimization*, 8(2):324–364.

Neumann, J. v. (1928). Zur Theorie der Gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320.

Ponsen, M. J. V., de Jong, S., and Lanctot, M. (2011). Computing approximate Nash equilibria and robust best-responses using sampling. *J. Artif. Intell. Res.*, 42:575–605.

Rakhlin, S. and Sridharan, K. (2013). Optimization, learning, and games with predictable sequences. In *Neural Information Processing Systems (NIPS)*.

Schneider, R. (2014). Convex bodies: The Brunn-Minkowski theory. *Encyclopedia of Mathematics and its Applications*, 1(151).

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. (2015). Fast convergence of regularized learning in games. In *Neural Information Processing Systems (NIPS)*.

Tammelin, O., Burch, N., Johanson, M., and Bowling, M. (2015). Solving heads-up limit Texas Hold'em. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

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

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. (2008). Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*.

A. General property of mirror descent

We first state a property of mirror descent which will be useful to prove Theorem 1.

Lemma 7. *Let $p \geq 1$ and $q \geq 1$ such that $1/p + 1/q = 1$. Let φ be a strongly convex function with respect to the ℓ_p-norm $\|\cdot\|_p$ with some modulus σ. Define π_{t+1} as*

$$\pi_{t+1} = \arg\max_{\pi \in \Delta} \left[\pi \cdot \delta_t - D_\varphi(\pi, \pi_t)\right], \tag{11}$$

Then we have, for any $\pi \in \Delta$,

$$D_\varphi(\pi, \pi_{t+1}) \leq D_\varphi(\pi, \pi_t) + (\pi_t - \pi) \cdot \delta_t + (2/\sigma)\|\delta_t\|_q^2.$$

Proof of Lemma 7. Let $J(\pi) \stackrel{\text{def}}{=} -\pi \cdot \delta_t + D_\varphi(\pi, \pi_t)$. J is differentiable and since π_{t+1} minimizes J over Δ, the gradient $\nabla J(\pi_{t+1}) = -\delta_t + \nabla\varphi(\pi_{t+1}) - \nabla\varphi(\pi_t)$ forms an acute angle with $(\pi - \pi_{t+1})$, for all $\pi \in \Delta$. Thus

$$\Big[-\delta_t + \nabla\varphi(\pi_{t+1}) - \nabla\varphi(\pi_t)\Big] \cdot (\pi - \pi_{t+1}) \geq 0.$$

From this inequality and from the definition of π_{t+1} we deduce

$$\begin{aligned}
D_\varphi(\pi, \pi_{t+1}) &= \varphi(\pi) - \varphi(\pi_{t+1}) - \nabla\varphi(\pi_{t+1}) \cdot (\pi - \pi_{t+1}) \\
&\leq \varphi(\pi) - \varphi(\pi_{t+1}) - \big[\delta_t + \nabla\varphi(\pi_t)\big] \cdot (\pi - \pi_{t+1}) \\
&= D_\varphi(\pi, \pi_t) + \varphi(\pi_t) - \varphi(\pi_{t+1}) - \delta_t \cdot (\pi - \pi_{t+1}) - \nabla\varphi(\pi_t) \cdot (\pi_t - \pi_{t+1}) \\
&= D_\varphi(\pi, \pi_t) - D_\varphi(\pi_{t+1}, \pi_t) - \delta_t \cdot (\pi - \pi_{t+1}) \\
&= D_\varphi(\pi, \pi_t) + \delta_t \cdot (\pi_t - \pi) + \delta_t \cdot (\pi_{t+1} - \pi_t) - D_\varphi(\pi_{t+1}, \pi_t).
\end{aligned}$$

Now, noticing that $J(\pi_{t+1}) \leq J(\pi_t)$, we have

$$-\pi_{t+1} \cdot \delta_t + D_\varphi(\pi_{t+1}, \pi_t) \leq -\pi_t \cdot \delta_t,$$

thus we deduce:

$$\begin{aligned}
\|\pi_{t+1} - \pi_t\|_p \|\delta_t\|_q &\geq (\pi_{t+1} - \pi_t) \cdot \delta_t \\
&\geq D_\varphi(\pi_{t+1}, \pi_t) \\
&\geq (\sigma/2)\|\pi_{t+1} - \pi_t\|_p^2,
\end{aligned} \tag{12}$$

where the first inequality is Hölder's inequality and the last inequality follows from the strong convexity of h. We deduce that $\|\pi_{t+1} - \pi_t\|_p \leq (2/\sigma)\|\delta_t\|_q$, thus

$$\begin{aligned}
D_\varphi(\pi, \pi_{t+1}) &\leq D_\varphi(\pi, \pi_t) + \delta_t \cdot (\pi_t - \pi) + (2/\sigma)\|\delta_t\|_q^2 - D_\varphi(\pi_{t+1}, \pi_t) \\
&\leq D_\varphi(\pi, \pi_t) + \delta_t \cdot (\pi_t - \pi) + (2/\sigma)\|\delta_t\|_q^2.
\end{aligned}$$

□

B. Proof of Theorem 1

Proof of Theorem 1. Let us use Lemma 7 for each player i with the choice $\delta_t = \eta_t Q_i^{\bar{\pi}_{-i,t}}$ and the policy π_i^*. We have

$$\begin{aligned}
J_{\pi^*}(t+1) &= \sum_{i \in \{1,2\}} D_\varphi(\pi_i^*, \pi_{i,t+1}) \\
&\leq \sum_{i \in \{1,2\}} \Big[D_\varphi(\pi_i^*, \pi_{i,t}) + \eta_t(\pi_{i,t} - \pi_i^*) \cdot Q_i^{\bar{\pi}_{-i,t}} + 2/\sigma\eta_t^2 \|Q_i^{\bar{\pi}_{-i,t}}\|_q^2\Big] \\
&\leq J_{\pi^*}(t) + \eta_t \sum_{i \in \{1,2\}} (\pi_{i,t} - \pi_i^*) \cdot Q_i^{\bar{\pi}_{-i,t}} + c\eta_t^2,
\end{aligned}$$

with $c = 4|A|^{1/q}Q_{\max}^2/\sigma$. Now we have

$$\begin{aligned}
\sum_{i\in\{1,2\}} (\pi_{i,t} - \pi_i^*)\cdot Q_i^{\bar{\pi}_{-i,t}} &= \sum_{i\in\{1,2\}} V_i^{(\pi_{i,t},\bar{\pi}_{-i,t})} - V_i^{(\pi_i^*,\bar{\pi}_{-i,t})} \\
&\overset{(i)}{=} -\sum_{i\in\{1,2\}} V_i^{(\bar{\pi}_{i,t},\pi_{-i,t})} + V_i^{(\bar{\pi}_i,\pi_{-i,t}^*)} \\
&\overset{(ii)}{=} \underbrace{-\sum_{i\in\{1,2\}} \left(V_i^{(\bar{\pi}_{i,t},\pi_{-i,t})} - V_i^{(\pi_{i,t},\pi_{-i,t})}\right)}_{=I(\bar{\pi}_t,\pi_t)} + \sum_{i\in\{1,2\}} \underbrace{\left(V_i^{(\bar{\pi}_i,\pi_{-i,t}^*)} - V_i^{(\pi_i^*,\pi_{-i,t}^*)}\right)}_{\leq 0 \text{ from Nash}} \\
&\leq -I(\bar{\pi}_t,\pi_t),
\end{aligned}$$

where we used the property that the game is zero-sum (i.e., $V_i^{\pi} = -V_{-i}^{\pi}$ for any policy π) in (i) and (ii), as well as the definition of the improvement $I(\bar{\pi}_t,\pi_t)$ and the property of a Nash eq.

We deduce

$$J_{\pi^*}(t+1) \leq J_{\pi^*}(t) - \eta_t I(\bar{\pi}_t,\pi_t) + c\eta_t^2.$$

□

C. Proof of Lemma 1

Proof of Lemma 1. We have $I^*(\pi) = \max_{\pi_1',\pi_2'} V^{(\pi_1',\pi_2)} - V^{(\pi_1,\pi_2')}$. Define

$$\begin{aligned}
f_1(\pi_2) &\overset{\text{def}}{=} \max_{\pi_1'} V^{(\pi_1',\pi_2)} - \min_{\pi_2'}\max_{\pi_1'} V^{(\pi_1',\pi_2')}, \\
f_2(\pi_1) &\overset{\text{def}}{=} \max_{\pi_1'}\min_{\pi_2'} V^{(\pi_1',\pi_2')} - \min_{\pi_2'} V^{(\pi_1,\pi_2')}.
\end{aligned}$$

From the minimax theorem we have that $\min_{\pi_2'}\max_{\pi_1'} V^{(\pi_1',\pi_2')} = \max_{\pi_1'}\min_{\pi_2'} V^{(\pi_1',\pi_2')}$, thus $I^*(\pi) = f_1(\pi_2) + f_2(\pi_1)$.

Since the following reasoning will apply for both f_1 and f_2, let us define f to be either f_1 or f_2. f is the maximum of affine functions defined over the simplex, thus is convex. Now since each function is linear, the maximum is reached on (at least) one vertex of the simplex. Thus f is the maximum –over a *finite number* of functions– of affine functions. Thus f is piecewise affine defined on a partitioning of the simplex where each piece of this partition is an intersection of half spaces, thus a convex polytope. A convex polytope also has the property of being defined as the convex hull of a finite number of extreme points (called vertices).

f is non-negative, so let us denote by X_0 the set of points x such that $f(x) = 0$ (X_0 is the restriction to the set of a Player's policies of the set of Nash equilibria, Π^*). X_0 is also a (closed) convex polytope.

Let V denote the (finite) set of vertices of all the polytopes. For any x in the simplex, write $P(x)$ the ℓ_2-projection of x onto X_0. We define

$$\kappa = \min_{v\in V, f(v)>0} \frac{f(v) - f(P(v))}{\|v - P(v)\|_2}.$$

Since V is a finite set, the minimum is reached for some value $\kappa > 0$.

Now consider any point x outside of X_0 and its ℓ_2-projection $x_0 = P(x)$ onto X_0. x belongs to some convex polytope with corresponding vertices v_i, thus $x = \sum_i \lambda_i v_i$, for some non-negative coefficients λ_i which sum to 1 (the so-called barycentric coordinates). Let us introduce $x_0' = \sum_i \lambda_i P(v_i)$ the barycenter of the projections of the vertices. Since X_0 is convex, then $x_0' \in X_0$.

We have

$$\begin{aligned}\frac{f(x)-f(P(x))}{\|x-P(x)\|_2} &\overset{(a)}{\geq} \frac{f(x)-f(x')}{\|x-x'\|_2}\\ &\overset{(b)}{=} \frac{\sum_i \lambda_i\big[f(v_i)-f(P(v_i))\big]}{\|\sum_i \lambda_i(v_i-P(v_i))\|_2}\\ &\overset{(c)}{\geq} \frac{\sum_i \lambda_i\big[f(v_i)-f(P(v_i))\big]}{\sum_i \lambda_i\|v_i-P(v_i)\|_2}\\ &\overset{(d)}{\geq} \kappa,\end{aligned}$$

where (a) results from $\|x-P(x)\|_2 \leq \|x-x'\|_2$ since $P(x)$ is the projection of x onto X_0, (b) from the linearity of f in the polytope, (c) from the convexity of the ℓ_2-norm, (d) from the definition of κ thus that for each i, $\frac{f(v_i)-f(P(v_i))}{\|v_i-P(v_i)\|_2} \geq \kappa$. We deduce that

$$f(x) = f(x) - f(P(x)) \geq \kappa\|x-P(x)\|_2 = \min_{y\in X_0}\|x-y\|_2.$$

This argument, applied successively to f_1 and f_2, proves that there exists κ_1 and κ_2 such that

$$\begin{aligned}I^*(\pi) &= f_1(\pi_2)+f_2(\pi_1)\\ &\geq \min_{(\pi_1',\pi_2')\in\Pi^*} \kappa_1\|\pi_2-\pi_2'\|_2+\kappa_2\|\pi_1-\pi_1'\|_2\\ &\geq \kappa \min_{\pi'\in\Pi^*}\|\pi-\pi'\|_2,\end{aligned}$$

with $\kappa = \min(\kappa_1,\kappa_2)$. □

D. Lower bound on κ

In this section, we give an interpretable lower bound on the value of the game-dependent parameter κ. This analysis contrasts with the exact characterization in (Mordukhovich et al., 2010). While we do not believe the bound to be tight, it provides some intuition as to what characteristics of the game affect κ.

Lemma 8. *Let R be the payoff matrix specifying the zero-sum game, and let P be the orthogonal projection operator onto the subspace of vectors with coordinates summing to zero. Then we have*

$$\kappa \geq \min\left(\min_{\substack{w\in\mathbb{R}^{A_1}_{\geq 0}\\ w\notin\mathrm{Ker}(PR^\top)}} \max_i \frac{e_i^\top RP^\top PR^\top w}{\|PR^\top w\|_2}, \min_{\substack{w\in\mathbb{R}^{A_2}_{\geq 0}\\ w\notin\mathrm{Ker}(PR)}} \max_i \frac{e_i^\top R^\top P^\top PRw}{\|PRw\|_2}\right).$$

Proof of Lemma 8. Recalling the notation in the proof of Lemma 1, we have

$$\kappa_1 = \min_{v\notin X_0}\frac{f(v)-f(P(v))}{\|v-P(v)\|},$$

where X_0 is the set of strategies for player 2 that feature in Nash equilibria, and $f=f_1$ returns the exploitability of its input policy. The numerator of the function on the right hand-side is convex and polyhedral (in v), and minimized on X_0 with value 0, so the minimum of this expression is attained by v arbitrarily close to X_0. Thus, we can restrict ourselves to consider pairs of variables $(v,P(v))$, where $v\in P(v)+N_{X_0}(P(v))$, where

$$N_{X_0}(P(v)) = \left\{\nu\in\mathbb{R}^{A_2}\middle|\sum_a \nu(a)=0\,,\langle P(v)+\nu,x\rangle\leq 0\ \ \forall x\in X_0\right\}\setminus\{0\},$$

is the normal cone to X_0 at $P(v)$, restricted to the linear subspace $\{x\in\mathbb{R}^{A_2}|\sum_i x_i=0\}$, and $\|v-P(v)\|$ is arbitrarily small. Now note that the linear constraints defining X_0 have a straightforward form:

$$\pi_2\in X_0 \iff \delta^\top R\pi_2\leq\Lambda \text{ for all deterministic policies } \delta\,, \tag{13}$$

where Λ is the value of the game. We may now use results from convex analysis to write down the form of $N_{X_0}(P(v))$; let I be the set of row indices of R for which $e_i^\top RP(v) = \Lambda \iff i \in I$. Then we have (see e.g. Theorem 2.4.9 of Schneider (2014)):

$$N_{X_0}(P(v)) = \left\{ \sum_{i \in I} \lambda_i PR^\top e_i \;\middle|\; \lambda_i \geq 0 \;\; \forall i \in I \right\} \setminus \{0\},$$

where P is orthogonal projection into $\{x \in \mathbb{R}^{A_2} | \sum_i x_i = 0\}$. Further, for vectors $n \in N_{X_0}(P(v))$ sufficiently small, we have $\text{BR}(P(v)+n) \subseteq \text{BR}(P(v)) = \{\sum_{i \in I} \alpha_i e_i | (\alpha_i)_{i \in I} \in \Delta_I\}$. We can therefore calculate, for such a $v = P(v) + n$, as follows:

$$\begin{aligned} \frac{f(v) - f(P(v))}{\|v - P(v)\|} &= \frac{\max_i e_i^\top R[P(v)+n] - \max_i e_i^\top RP(v)}{\|n\|} \\ &= \frac{\max_i e_i^\top Rn}{\|n\|} \\ &= \max_i e_i^\top R \frac{n}{\|n\|}. \end{aligned} \tag{14}$$

Recall that each normal vector n must take the form $PR^\top w$, for some non-zero $w \in \mathbb{R}^{A_1}_{\geq 0}$. We can further restrict to $w \notin \text{Ker}(PR^\top)$, since we consider only non-zero normal vectors n. There are additional restrictions we can make as to which components of w are non-zero, since they must arise from hyperplanes defining X_0 that have a common intersection point on the boundary of the polytope. Pursuing this direction would complicate the analysis, but result in a tighter bound on κ. We do not make these additional restrictions here, and note that minimizing Expression (14) over a larger class of vectors n still yields a valid lower bound on κ. Using this parametrisation, we have

$$\kappa_1 \geq \min_{\substack{w \in \mathbb{R}^{A_1}_{\geq 0} \\ w \notin \text{Ker}(PR^\top)}} \max_i \frac{e_i^\top RPR^\top w}{\|PR^\top w\|}.$$

Using the fact that $P^\top = P$ and $P^2 = P$, we can write

$$\kappa_1 \geq \min_{\substack{w \in \mathbb{R}^{A_1}_{\geq 0} \\ w \notin \text{Ker}(PR^\top)}} \max_i \frac{e_i^\top RP^\top PR^\top w}{\|PR^\top w\|}.$$

Repeating the same analysis with the roles of the players reversed, and overloading P to also represent the equivalent orthogonal projection in $\mathbb{R}^{A_1}$ yields

$$\kappa_2 \geq \min_{\substack{w \in \mathbb{R}^{A_2}_{\geq 0} \\ w \notin \text{Ker}(PR)}} \max_i \frac{e_i^\top R^\top P^\top PRw}{\|PRw\|}.$$

So, we obtain the result

$$\kappa \geq \min \left(\min_{\substack{w \in \mathbb{R}^{A_1}_{\geq 0} \\ w \notin \text{Ker}(PR^\top)}} \max_i \frac{e_i^\top RP^\top PR^\top w}{\|PR^\top w\|}, \min_{\substack{w \in \mathbb{R}^{A_2}_{\geq 0} \\ w \notin \text{Ker}(PR)}} \max_i \frac{e_i^\top R^\top P^\top PRw}{\|PRw\|} \right)$$

□

E. A simple matrix game

Consider the game in normal form defined by the reward matrix $R = \begin{pmatrix} 0 & 1-\varepsilon \\ 2 & 1 \end{pmatrix}$, for some $\varepsilon \in (0,1)$. The Nash eq. is $\pi_1^* = (0,1)$, $\pi_2^* = (0,1)$ and the value of the game is 1. Let us write a policy profile $\pi = (p,q)$ with $p, q \in [0,1]$, where

$\pi_1 = (p, 1-p)$ and $\pi_2 = (q, 1-q)$. The value of any policy $\pi = (p, q)$ is

$$V^{\pi} = V^{(p,q)} = (1-\varepsilon)p(1-q) + 2(1-p)q + (1-p)(1-q) = -\varepsilon p(1-q) + q - 2pq + 1.$$

We have

$$\max_p V^{p,q} = V^{p=0,q} = q+1, \text{ and } \min_q V^{p,q} = V^{p,q=0} = -\varepsilon p + 1.$$

We deduce that the exploitability of the policy $\pi = (p, q)$ is

$$\begin{aligned} I^*(\pi) &= \max_p V^{p,q} - \min_q V^{p,q} \\ &= q + \varepsilon p \\ &\geq \varepsilon\sqrt{p^2+q^2} \\ &= \kappa\sqrt{\|\pi_1 - \pi_1^*\|_2^2 + \|\pi_2 - \pi_2^*\|_2^2} \\ &= \kappa\|\pi - \pi^*\|_2, \end{aligned}$$

for $\kappa = \varepsilon/\sqrt{2}$.

F. Proof of Theorem 3

Proof. A first property is that if $\mu_i^{(\pi_i,\bar{\pi}_{-i})}(x) \neq 0$ then we have:

$$\begin{aligned} \sum_{h\in x} \mu_{\neq i}^{\bar{\pi}}(h) Q_i^{(\pi_i,\bar{\pi}_{-i})}(h,a) &= \sum_{h\in x} \frac{\mu^{(\pi_i,\bar{\pi}_{-i})}(h)}{\mu_i^{\pi}(h)} Q_i^{(\pi_i,\bar{\pi}_{-i})}(h,a) \\ &= \frac{1}{\mu_i^{\pi}(x)} \sum_{h\in x} \mu^{(\pi_i,\bar{\pi}_{-i})}(h) Q_i^{(\pi_i,\bar{\pi}_{-i})}(h,a) \\ &= \frac{\mu^{(\pi_i,\bar{\pi}_{-i})}(x)}{\mu_i^{\pi}(x)} Q_i^{(\pi_i,\bar{\pi}_{-i})}(x,a) \\ &= \mu_{\neq i}^{\bar{\pi}}(x) Q_i^{(\pi_i,\bar{\pi}_{-i})}(x,a). \end{aligned} \tag{15}$$

Thus in the case $\mu_i^{(\pi_i,\bar{\pi}_{-i})}(x) \neq 0$ the MAIO-IIG algorithm can rewrite:

$$\pi_{i,t+1}(x) \in \operatorname*{arg\,min}_{\pi_i\in\Delta(A_i)} \Big[D_{\varphi}(\pi_i, \pi_{i,t}(x)) - \eta_t \sum_{a\in A_i} \pi_i(a) \mu_{\neq i}^{\bar{\pi}_t}(x) Q_i^{(\pi_{i,t},\bar{\pi}_{-i,t})}(x,a) \Big].$$

For simplicity of the notation, let us drop the time index for the policies, writing π_i, $\bar{\pi}$ and π_{-i} instead $\pi_{i,t}$, $\bar{\pi}_t$ and $\pi_{-i,t}$, respectively. Defining

$$\delta_{i,t}(x,a) \stackrel{\text{def}}{=} \eta_t \sum_{h\in x} \mu_{\neq i}^{\bar{\pi}}(h) Q_i^{(\pi_i,\bar{\pi}_{-i})}(h,a),$$

from Lemma 7, we have

$$\begin{aligned} J_{\pi^*}(t+1) &= \sum_{i\in\{1,2\}} \sum_{x\in X_i} \mu_i^{\pi^*}(x) D_{\varphi}(\pi_i^*(x), \pi_{i,t+1}(x)) \\ &\leq \sum_{i\in\{1,2\}} \sum_{x\in X_i} \mu_i^{\pi^*}(x) \Big(D_{\varphi}(\pi_i^*(x), \pi_{i,t+1}(x)) + (\pi_{i,t}(x) - \pi_i^*(x)) \cdot \delta_{i,t} + 2/\sigma \|\delta_{i,t}(x,a)\|_q^2 \Big) \\ &= J_{\pi^*}(t) + \sum_{i\in\{1,2\}} \sum_{x\in X_i} \mu_i^{\pi^*}(x) \Big((\pi_{i,t}(x) - \pi_i^*(x)) \cdot \delta_{i,t}(x,a) + 2/\sigma \|\delta_{i,t}(x,a)\|_q^2 \Big) \\ &\leq J_{\pi^*}(t) + \eta_t \underbrace{\sum_{i\in\{1,2\}} \sum_{x\in X_i} \mu_i^{\pi^*}(x) \sum_{h\in x} \mu_{\neq i}^{\bar{\pi}}(h) \sum_a (\pi_{i,t}(a|x) - \pi_i^*(a|x)) Q_i^{(\pi_i,\bar{\pi}_{-i})}(h,a)}_{(a)} + c\eta_t^2, \end{aligned} \tag{16}$$

with $c \stackrel{\text{def}}{=} \frac{4}{\sigma}|A|^{2/q}Q_{\max}^2 L_{\max}$ and $Q_{\max} = \max_\pi \max_{h\in H, a\in A} |Q^\pi(h,a)|$ and $L_{\max} = \max_\pi \sum_x \mu^\pi(x)$.

Now using Property (8), and the fact that $\mu_i^\pi(h) = \mu_i^\pi(x)$ for any $h \in x \in X_i$, we have

$$\begin{aligned}\mu_i^{\pi^*}(x)\sum_{h\in x}\mu_{\neq i}^{\bar\pi}(h)Q_i^{(\pi_i,\bar\pi_{-i})}(h,a) &= \sum_{h\in x}\mu_i^{\pi^*}(h)\mu_{\neq i}^{\bar\pi}(h)Q_i^{(\pi_i,\bar\pi_{-i})}(h,a)\\ &= \sum_{h\in x}\mu^{(\pi_i^*,\bar\pi_{-i})}(h)Q_i^{(\pi_i,\bar\pi_{-i})}(h,a).\end{aligned}$$

Thus

$$(a) = \sum_{i\in\{1,2\}}\sum_{x\in X_i}\sum_{h\in x}\mu^{(\pi_i^*,\bar\pi_{-i})}(h)\sum_a(\pi_{i,t}(a|x)-\pi_i^*(a|x))Q_i^{(\pi_i,\bar\pi_{-i})}(h,a).$$

Finally using Lemma 9 below with the policies π, π^* and $\bar\pi$, and summing over $i\in\{1,2\}$, we deduce that

$$(a) = \sum_{i\in\{1,2\}}\left(V_i^{(\pi_i,\bar\pi_{-i})} - V_i^{(\pi_i^*,\bar\pi_{-i})}\right)$$

Now from the definion of the improvement, we deduce

$$\begin{aligned}(a) &= \sum_{i\in\{1,2\}}\left(V_i^{(\pi_i,\bar\pi_{-i})} - V_i^{(\pi_i^*,\bar\pi_{-i})}\right)\\ &= -\sum_{i\in\{1,2\}}\left(V_i^{(\bar\pi_i,\pi_{-i})} - V_i^{(\bar\pi_i,\pi_{-i}^*)}\right)\\ &= -\underbrace{\sum_{i\in\{1,2\}}\left(V_i^{(\bar\pi_i,\pi_{-i})} - V_i^{(\pi_i,\pi_{-i})}\right)}_{=I(\bar\pi,\pi)} + \sum_{i\in\{1,2\}}\Big(\underbrace{V_i^{(\bar\pi_i,\pi_{-i}^*)} - V_i^{(\pi_i^*,\pi_{-i}^*)}}_{\leq 0 \text{ from Nash}}\Big)\\ &\leq -I(\bar\pi,\pi).\end{aligned}$$

Using this in (16) we deduce

$$J_{\pi^*}(t+1) \leq J_{\pi^*}(t) - \eta_t I(\bar\pi_t,\pi_t) + c\eta_t^2.$$

□

G. Auxiliary results

Lemma 9. *For any triplet of policies π, π' and $\bar\pi$ and any Player i, we have*

$$V_i^{(\pi_i,\bar\pi_{-i})} - V_i^{(\pi_i',\bar\pi_{-i})} = \sum_{x\in X_i}\mu^{(\pi_i',\bar\pi_{-i})}(x)\sum_a(\pi_i(a|x)-\pi_i'(a|x))Q_i^{(\pi_i,\bar\pi_{-i})}(x,a).$$

Proof. For any Player i, let us define the policies $\tilde\pi_i$ and $\tilde\pi_i'$:

$$\tilde\pi_i(a|h) \stackrel{\text{def}}{=} \begin{cases}\pi_i(a|x(h)) & \text{if } h\in H_i\\ \bar\pi_i(a|x(h)) & \text{otherwise}\end{cases} \text{ and } \tilde\pi_i'(a|h) \stackrel{\text{def}}{=} \begin{cases}\pi_i'(a|x(h)) & \text{if } h\in H_i\\ \bar\pi_i(a|x(h)) & \text{otherwise}\end{cases},$$

Since $\tilde\pi_i$ and $\tilde\pi_i'$ coincide outside of H_i, we have

$$\begin{aligned}(a) &\stackrel{\text{def}}{=} \sum_{x\in X_i}\mu^{(\pi_i',\bar\pi_{-i})}(x)\sum_a(\pi_i(a|x)-\pi_i'(a|x))Q_i^{(\pi_i,\bar\pi_{-i})}(x,a)\\ &= \sum_{x\in X_i}\sum_{h\in x}\mu^{(\pi_i',\bar\pi_{-i})}(h)\sum_a(\pi_i(a|x)-\pi_i'(a|x))Q_i^{(\pi_i,\bar\pi_{-i})}(h,a)\\ &= \sum_{h\in H}\mu_i^{\tilde\pi'}(h)\sum_a(\tilde\pi_i(a|h)-\tilde\pi_i'(a|h))Q_i^{\tilde\pi_i}(h,a).\end{aligned}$$

From the definition of the value function $V_i^{\tilde{\pi}_i}(h) = \sum_a \tilde{\pi}_i(a|h) Q_i^{\tilde{\pi}_i}(h,a)$, and the Bellman equation for Q at the history level is:

$$Q_i^{\tilde{\pi}_i}(h,a) = r_i(h,a) + \sum_{h'} p(h'|h,a) V_i^{\tilde{\pi}_i}(h').$$

Thus

$$(a) \quad = \quad \sum_{h \in H} \mu_i^{\tilde{\pi}'}(h) \Big(V_i^{\tilde{\pi}_i}(h) - \sum_a \tilde{\pi}'_i(a|h) \big[r_i(h,a) + \sum_{h' \in H} p(h'|h,a) V_i^{\tilde{\pi}_i}(h') \big] \Big).$$

Now, from the balance equation (3) we have

$$\sum_{h \in H} \mu_i^{\tilde{\pi}'}(h) \sum_a \tilde{\pi}'_i(a|h) p(h'|h,a) = \mu_i^{\tilde{\pi}'}(h') - \rho_0(h').$$

Thus

$$\begin{aligned}
(a) \quad &= \quad \sum_{h \in H} \mu_i^{\tilde{\pi}'}(h) V_i^{\tilde{\pi}_i}(h) - \sum_{h' \in H} \Big(\mu_i^{\tilde{\pi}'}(h') - \rho_0(h') \Big) V_i^{\tilde{\pi}_i}(h') - \sum_{h \in H} \mu_i^{\tilde{\pi}'}(h) \sum_a \tilde{\pi}'_i(a|h) r_i(h,a) \\
&= \quad \sum_{h' \in H} \rho_0(h') V_i^{\tilde{\pi}_i}(h') - \sum_{h \in H} \mu_i^{\tilde{\pi}'}(h) \sum_a \tilde{\pi}'_i(a|h) r_i(h,a) \\
&= \quad V_i^{\tilde{\pi}_i} - V_i^{\tilde{\pi}'_i},
\end{aligned}$$

where we used (5) and (6) to derive the last equality. □

H. Exploitability and J_{π^*}-distance

Lemma 2. For any policy π, the exploitability of π is upper bounded by the ℓ_2-energy distance (as defined by $\min_{\pi^* \in \Pi^*} J_{\pi^*}(\pi)$) between π and Π^*:

$$I^*(\pi)^2 \quad \leq \quad L_{\max}^2 |A| Q_{\max}^2 \min_{\pi^* \in \Pi^*} J_{\pi^*}(\pi),$$

where $L_{\max} \stackrel{\text{def}}{=} \max_\pi \sum_{x \in X} \mu^\pi(x)$, $Q_{\max} \stackrel{\text{def}}{=} \max_\pi \max_{h,a} |Q^\pi(h,a)|$, and $J_{\pi^*}(\pi) \stackrel{\text{def}}{=} \sum_i \sum_{x \in X_i} \mu_i^{\pi^*}(x) \|\pi_i(x) - \pi_i^*(x)\|_2^2$.

Proof. For any Nash eq. $\pi^* \in \Pi^*$, from the definition of the exploitability, and since $I^*(\pi^*) = 0$, we have

$$I^*(\pi) = \max_{\pi'} \sum_i V_i^{(\pi'_i, \pi_{-i})} - \max_{\pi'_i} \sum_i V_i^{(\pi'_i, \pi^*_{-i})} \leq \max_{\pi'} \sum_i \Big(V_i^{(\pi'_i, \pi_{-i})} - V_i^{(\pi'_i, \pi^*_{-i})} \Big).$$

Since the game is zero-sum two-player,

$$\sum_i \Big(V_i^{(\pi'_i, \pi_{-i})} - V_i^{(\pi'_i, \pi^*_{-i})} \Big) = - \sum_i \Big(V_i^{(\pi_i, \pi'_{-i})} - V_i^{(\pi^*_i, \pi'_{-i})} \Big).$$

Thus

$$I^*(\pi) \leq \max_{\pi'} \sum_i \Big(V_i^{(\pi^*_i, \pi'_{-i})} - V_i^{(\pi_i, \pi'_{-i})} \Big). \tag{17}$$

Now, using Lemma 9, we have

$$\begin{aligned}\Big(\sum_i V_i^{(\pi_i^*,\pi'_{-i})} - V_i^{(\pi_i,\pi'_{-i})}\Big)^2 &= \Big(\sum_i \sum_{x\in X_i} \mu^{(\pi_i^*,\pi'_{-i})}(x) \sum_a (\pi_i^*(a|x) - \pi_i(a|x)) Q^{(\pi_i^*,\pi'_{-i})}(x,a)\Big)^2 \\ &\overset{(a)}{\leq} \Big(\sum_i \sum_{x\in X_i} \mu^{(\pi_i^*,\pi'_{-i})}(x)\Big)\Big[\sum_i \sum_{x\in X_i} \mu^{(\pi_i^*,\pi'_{-i})}(x) \\ &\quad \times \Big(\sum_a (\pi_i^*(a|x) - \pi_i(a|x)) Q^{(\pi_i^*,\pi'_{-i})}(x,a)\Big)^2\Big] \\ &\overset{(b)}{\leq} L_{\max}^2 |A| Q_{\max}^2 \sum_i \sum_{x\in X_i} \mu_i^{\pi^*}(x) \|\pi_i^*(x) - \pi_i(x)\|_2^2 \\ &= L_{\max}^2 |A| Q_{\max}^2 J_{\pi^*}(\pi),\end{aligned}$$

where we used Cauchy-Schwarz inequality for (a), i.e. $\sum_x f(x)g(x) \leq \big(\sum_x f(x)\big)^{1/2}\big(\sum_x f(x)g^2(x)\big)^{1/2}$, and in (b) we used the definition of $L_{\max}$, $Q_{\max}$ and the property that $\mu^{(\pi^*,\pi'_{-i})}(x) = \mu_i^{\pi^*}(x)\mu_{\neq i}^{\pi'}(x) \leq \mu_i^{\pi^*}(x)$. Since this inequality is true for all π^*, we deduce from (17) that

$$I^*(\pi)^2 \leq L_{\max}^2 |A| Q_{\max}^2 \min_{\pi^*\in\Pi^*} J_{\pi^*}(\pi).$$

□

An immediate consequence of this result is that convergence in policy (i.e., $\min_{\pi^*\in\Pi^*} J_{\pi^*}(\pi_t) \to 0$) implies convergence in value (i.e., $I(\pi^*, \pi_t) \to 0$) since $I(\pi^*, \pi_t) \leq I^*(\pi_t)$.

I. The MDP $\mathcal{M}_i^{\pi_{-i}}$

For any Player i, fix the opponent's policy π_i and consider the observation process $(x_k = x(h_k))_{k\geq 0: i(h_k)=i}$ at successive times when it it Player i's turn to play.

In a IIG, this process is in general a Partiable Observation Markov Decision Process (POMDP). Indeed the probability of the next observation x_{k+1} depends not only on the current observation x_k, action a_k chosen by Player i, but also on what Player i played to reach information node x_k.

However, under the perfect recall assumption, we have the remarkable property that this observation process is actually a Markov Decision Process (MDP), which we write as $\mathcal{M}_i^{\pi_{-i}}$.

Proposition 1. *Consider a fixed opponent π_{-i}. Under perfect recall (Assumption 1), the transitions and rewards between observations for player i form a MDP, writen $\mathcal{M}_i^{\pi_{-i}}$.*

Proof. Consider two consecutive observations $x, x' \in X_i$ for Player i. This means that for histories $h \in x$ and $h' \in x'$ such that h is an ancestor of h', there exists a path $(h_0 = h, a_0 = a, h_1, a_1, \ldots, h_{n-1}, a_{n-1}, h_n = h')$ from h to h', for some $n > 0$ (if the Players play alternatively then $n = 2$), such that $h_j \in H_{-i}$ for all $j = 1 \ldots n-1$. Thus the transition probability from h to h' conditioned on a depends on π_{-i} only:

$$p^{\pi_{-i}}(h'|h,a) \overset{\text{def}}{=} p(h_1|h,a) \prod_{j=1}^{n-1} \pi(a_j|h_j) p(h_{j+1}|h_j, a_j).$$

Thus the probability of transitioning from x to x' given a is

$$\begin{aligned}p_i^{\pi_{-i}}(x'|x,a) &\overset{\text{def}}{=} \frac{\sum_{h'\in x'} \sum_{h\in x} \mu^\pi(h) p^{\pi_{-i}}(h'|h,a)}{\sum_{h\in x} \mu^\pi(h)} \\ &= \frac{\sum_{h'\in x'} \sum_{h\in x} \mu_{\neq i}^\pi(h) p^{\pi_{-i}}(h'|h,a)}{\sum_{h\in x} \mu_{\neq i}^\pi(h)},\end{aligned}$$

where the last equality comes from the fact that $\mu^\pi(h) = \mu_i^\pi(h)\mu_{\neq i}^\pi(h)$ and that $\mu_i^\pi(h)$ does not depend on the specific $h \in x$ but on x only. Thus $p_i^{\pi_{-i}}(x'|x,a)$ does not depend on π_i.

We can define similarly the reward obtained by player i along the path from x, a to x' as

$$\begin{aligned} r_i^{\pi_{-i}}(x,a) &\stackrel{\text{def}}{=} \sum_{h\in x} \frac{\mu^\pi(h)}{\sum_{h\in x}\mu^\pi(h)} \sum_{h'\in H_i} p^{\pi_{-i}}(h'|h,a) \sum_{j=0}^{n-1} r_i(h_j,a_j), \\ &= \sum_{h\in x} \frac{\mu_{\neq i}^\pi(h)}{\sum_{h\in x}\mu_{\neq i}^\pi(h)} \sum_{h'\in H_i} p^{\pi_{-i}}(h'|h,a) \sum_{j=0}^{n-1} r_i(h_j,a_j), \end{aligned}$$

which does not depend on π_i. The MDP $\mathcal{M}_i^{\pi_{-i}}$ is defined by the transition probabilities $p_i^{\pi_{-i}}(x'|x,a)$ and rewards $r_i^{\pi_{-i}}(x,a)$ which do not depend on the Player i's own policy. □

Thus it is interesting to notice that under the perfect recall Assumption 1, from the point of view of any Player i (and fixing the opponent's policy π_{-i}), the transitions between consecutive observations and corresponding rewards are Markovian, i.e., $p_i^{\pi_{-i}}(x'|x,a)$ and $r_i^{\pi_{-i}}(x,a)$ do not depend on Player i's own policy.

Also we can define the Bellman equations in $\mathcal{M}_i^{\pi_{-i}}$: The Q-values defined by (7) satisfy the Bellman equation for MDP $\mathcal{M}_i^{\pi_{-i}}$: for any $x \in X_i, a \in A$,

$$\begin{aligned} Q_i^\pi(x,a) &= r_i^{\pi_{-i}}(x,a) + \sum_{x'\in X_i} p_i^{\pi_{-i}}(x'|x,a) V_i^\pi(x') \\ V_i^\pi(x) &= \sum_a \pi(a|x) Q_i^\pi(x,a). \end{aligned}$$

J. Proof of Lemma 3

Lemma 3. There exists a two-player zero-sum imperfect information game such that there exists no $\kappa > 0$ such that for all π, $I^*(\pi) \geq \kappa \min_{\pi^*\in\Pi^*} \sqrt{J_{\pi^*}(\pi)}$.

Proof. Consider the following game:

- There are 5 possible histories: $h_0, \dots h_4$. Root is h_0. Player 1 plays in h_0 and h_1, player 2 plays in the other ones.
- There are 3 information nodes: $x_0 = \{h_0\}$, $x_1 = \{h_1\}$ for player 1, and $x_2 = \{h_2, h_3, h_4\}$ for player 2.
- Each player has 2 actions a_0, a_1. Transitions are deterministic: $(h_0, a_0 \to h_1)$, $(h_0, a_1 \to h_4)$, $(h_1, a_0 \to h_2)$, $(h_1, a_1 \to h_3)$. All other histories transition to a terminal state.
- Rewards are: $r_1(h_2, a_0) = 0$, $r_1(h_2, a_1) = 0$, $r_1(h_3, a_0) = 2$, $r_1(h_3, a_1) = 1$, $r_1(h_4, a_0) = 2$, $r_1(h_4, a_1) = 1$. All other rewards are 0.

Write any policy as $\pi = (p_0, p_1, q)$ where $p_0 = \pi(a_0|x_0)$, $p_1 = \pi(a_0|x_1)$, $q = \pi(a_0|x_2)$.

Notice that this problem is basically the same as the 2×2 matrix game defined above (with $\varepsilon = 1$) where the probability p is now replaced by the product $p_0 p_1$.

Thus the set of Nash equilibria (p_0^*, p_1^*, q^*) is defined by $p_0^* p_1^* = 0$ and $q^* = 0$ and the value of the game is 1. Notice that the set of Nash equilibria

$$\Pi^* = \{(p_0^* = 0, p_1^*, q^* = 0)\} \cup \{(p_0, p_1^* = 0, q^* = 0)\}$$

is a non-convex subset of the set of policies π, in contrary to the case of normal form games. This gives us a clue to why it is not possible to generalize the result of Lemma 1 to imperfect information games.

Let us consider the value of player 1 and write V for V_1. The value of any policy is:

$$V^{p_0,p_1,q} = -p_0 p_1 (1-q) + q - 2 p_0 p_1 q + 1.$$

Thus we have $\max_{p_0,p_1} V^{p_0,p_1,q} = 1+q$ and $\min_q V^{p_0,p_1,q} = 1-p_0p_1$. Thus the exploitability of any policy $\pi = (p_0, p_1, q)$ is

$$I^*(\pi) = \max_{p_0,p_1} V^{p_0,p_1,q} - \min_q V^{p_0,p_1,q} = q + p_0p_1.$$

Now the $J_{\pi^*}(\pi)$ distance of $\pi = (p_0, p_1, q)$ to a Nash $\pi^* = (p_0^*, p_1^*, q^*)$ is

$$\begin{aligned} J_{\pi^*}(\pi) &= \mu_1^{\pi^*}(x_0)\|\pi(x_0) - \pi^*(x_0)\|^2 + \mu_1^{\pi^*}(x_1)\|\pi(x_1) - \pi^*(x_1)\|^2 + \mu_2^{\pi^*}(x_2)\|\pi(x_2) - \pi^*(x_2)\|^2 \\ &= \|\pi(x_0) - \pi^*(x_0)\|^2 + p_0^*\|\pi(x_1) - \pi^*(x_1)\|^2 + \|\pi(x_2) - \pi^*(x_2)\|^2 \\ &= 2\big[(p_0 - p_0^*)^2 + p_0^*(p_1 - p_1^*)^2 + (q - q^*)^2\big]. \end{aligned}$$

Thus

$$\begin{aligned} \min_{\pi^* \in \Pi^*} J_{\pi^*}(\pi) &= \min\Big(\min_{\pi^*=(p_0^*=0,p_1^*,q^*=0)} J_{\pi^*}(\pi), \min_{\pi^*=(p_0^*,p_1^*=0,q^*=0)} J_{\pi^*}(\pi)\Big) \\ &= 2\min\Big(p_0^2 + q^2, \min_{\pi^*=(p_0^*,p_1^*=0,q^*=0)} (p_0 - p_0^*)^2 + p_0^*p_1^2 + q^2\Big). \end{aligned}$$

The mimimum is reached for $p_0^* = p_0 - p_1^2/2$ if $p_1^2 \le 2p_0$, and the corresponding value is

$$\min_{\pi^* \in \Pi^*} J_{\pi^*}(\pi) = 2\min\Big(p_0^2 + q^2, -p_1^4/4 + p_0p_1^2 + q^2\Big).$$

It is clear that there is no constant $\kappa > 0$ such that for all $p_0, p_1, q \in [0, 1]$ we have $I^*(\pi) \ge \kappa \min_{\pi^* \in \Pi^*} \sqrt{J_{\pi^*}(\pi)}$. For example choosing $p_0 = p_1 = \sqrt{q}$, this would mean there exists $\kappa > 0$ such that

$$q + q \ge 2\kappa\sqrt{\min\big(q + q^2, -q^2/4 + q^{3/2} + q^2\big)},$$

which is impossible for small enough q. □

K. Proof of Lemma 5

Lemma 5. For any π^* there exists two constant $\delta, c > 0$ such for any π such that $d(\pi^*, \pi) \le \delta$, we have

$$J_{\pi^*}(\pi) \le c\, d(\pi^*, \pi).$$

Proof. Let

$$\varepsilon \stackrel{\text{def}}{=} \min_{\{i,x,a:\mu_i^{\pi^*}(x)>0\}} \mu_i^{\pi^*}(x). \tag{18}$$

From its definition, $\varepsilon > 0$. Now define $\delta \stackrel{\text{def}}{=} \frac{\varepsilon^2}{4|A|}$, and assume that $d(\pi^*, \pi) \le \delta$. Define $e(x, a) \stackrel{\text{def}}{=} |\mu_i^{\pi^*}(a|x) - \mu_i^{\pi}(a|x)|$ for $i = I(x)$. Thus we have

$$\sum_a e(x, a) \le \sqrt{|A| \sum_a |\mu_i^{\pi^*}(a|x) - \mu_i^{\pi}(a|x)|^2} \le \sqrt{|A| d(\pi^*, \pi)} \le \frac{\varepsilon}{2} \le \frac{\mu_i^{\pi^*}(x)}{2}. \tag{19}$$

Now, for $x \in X_i$ such as $\mu_i^{\pi^*}(x) > 0$, we have

$$\begin{array}{rcccl} \frac{\mu_i^{\pi}(x,a)}{\mu_i^{\pi}(x)} - \frac{\mu^{\pi_i^*}(x,a)}{\mu^{\pi_i^*}(x)} & = & \pi_i(a|x) - \pi_i^*(a|x) & = & \frac{\mu_i^{\pi}(x,a)}{\mu_i^{\pi}(x)} - \frac{\mu^{\pi_i^*}(x,a)}{\mu^{\pi_i^*}(x)} \\ \frac{\mu_i^{\pi^*}(x,a) - e(x,a)}{\mu_i^{\pi^*}(x) + \sum_{a'} e(x,a')} - \frac{\mu^{\pi_i^*}(x,a)}{\mu^{\pi_i^*}(x)} & \le & & \stackrel{(i)}{\le} & \frac{\mu_i^{\pi^*}(x,a) + e(x,a)}{\mu_i^{\pi^*}(x) - \sum_{a'} e(x,a')} - \frac{\mu^{\pi_i^*}(x,a)}{\mu^{\pi_i^*}(x)} \\ -\frac{2e(x,a)}{\mu^{\pi_i^*}(x) + \sum_{a'} e(x,a')} & \le & & \le & \frac{2e(x,a)}{\mu^{\pi_i^*}(x) - \sum_{a'} e(x,a')} \\ -2\frac{e(x,a)}{\mu^{\pi_i^*}(x)} & \le & & \stackrel{(ii)}{\le} & 4\frac{e(x,a)}{\mu^{\pi_i^*}(x)}, \end{array}$$

where we used (19) in (i) and (ii). We deduce that

$$\begin{aligned} J_{\pi^*}(\pi) &= \sum_i \sum_{x \in X_i} \mu_i^{\pi^*}(x) \|\pi_i(x) - \pi_i^*(x)\|^2 \\ &\leq 16 \sum_i \sum_{x \in X_i} \mu_i^{\pi^*}(x) \sum_{a \in A} \frac{e(x,a)^2}{\mu_i^{\pi^*}(x)^2} \\ &= \frac{16}{\varepsilon} \sum_i \sum_{x \in X_i} \sum_{a \in A} \left| \mu_i^{\pi^*}(a|x) - \mu_i^{\pi}(a|x) \right|^2 \\ &= c\, d(\pi^*, \pi). \end{aligned}$$

with $c \stackrel{\text{def}}{=} 16/\varepsilon$. □

L. Proof of Theorem 4

Theorem 4 (Convergence of MAIO-BR). The sequence of policies produced by MAIO-BR algorithm with $\eta_t = I^*(\pi_t)/(2c)$ converges to the set of Nash equilibria, in the sense that $\lim_{t\to\infty} \min_{\pi^* \in \Pi^*} J_{\pi^*}(\pi_t) = 0$. Notice that from Lemma 2 we also deduce the result in exploitability: $\lim_{t\to\infty} I^*(\pi_t) = 0$.

Proof. The sequence of policies (π_t) lies in the compact space of policies, thus from the Bolzano-Weierstrass theorem, there exists (at least) one accumulation point to this sequence. Let us write $\pi^\#$ a such point and $(\pi_{\tau_t})_t$ a subsequence which converges to $\pi^\#$. Now we can apply Bolzano-Weierstrass theorem again to the subsequence of best responses $(\bar{\pi}_{\tau_t})_t$, so there exists a sub-subsequence $(\tau'_t)_t$ of the subsequence $(\tau_t)_t$ such that $(\bar{\pi}_{\tau'_t})_t$ converges to some policy, which we write $\bar{\pi}^\#$. First, let us prove that $\bar{\pi}^\#$ is a best response to $\pi^\#$. From the definition of $\bar{\pi}_{\tau'_t}$ being a best response to $\pi_{\tau'_t}$, we have that, for each player i,

$$V_i^{(\bar{\pi}_{i,\tau'_t}, \pi_{-i,\tau'_t})} \geq V_i^{(\pi'_i, \pi_{-i,\tau'_t})}, \text{ for any } \pi'.$$

Since the map $\pi \mapsto V_i^\pi$ is continuous, we can take this inequality to the limit $t \to \infty$ and deduce that

$$V_i^{(\bar{\pi}_i^\#, \pi_{-i}^\#)} \geq V_i^{(\pi'_i, \pi_{-i}^\#)}, \text{ for any } \pi',$$

thus $\bar{\pi}^\#$ is a best response to $\pi^\#$ and we have $I^*(\pi^\#) = I^*(\bar{\pi}^\#, \pi^\#)$.

Now, from Lemma 3 we have that for any $\pi^* \in \Pi^*$,

$$J_{\pi^*}(\pi_{t+1}) \leq J_{\pi^*}(\pi_t) - \frac{I^*(\pi_t)}{4c}. \tag{20}$$

Thus the sequence $(J_{\pi^*}(\pi_t))_t$ is non-increasing and lower bounded by 0, thus converges to some value $J_{\pi^*} \geq 0$.

The maps $\pi_1, \pi_2 \mapsto I(\pi_1, \pi_2)$ and $\pi \mapsto J_{\pi^\#}(\pi)$ being continuous, we can take the limit in the following inequality:

$$J_{\pi^*}(\pi_{\tau'_{t+1}}) \leq J_{\pi^*}(\pi_{\tau'_t+1}) \leq J_{\pi^*}(\pi_{\tau'_t}) - \frac{I(\bar{\pi}_{\tau'_t}, \pi_{\tau'_t})}{4c},$$

and deduce that $J_{\pi^*} \leq J_{\pi^*} - \frac{I(\bar{\pi}^\#, \pi^\#)}{4c}$, thus $I(\bar{\pi}^\#, \pi^\#) = I^*(\pi^\#) = 0$. Thus $\pi^\#$ is a Nash equilibrium.

Thus applying (20) with the Nash equilibrium $\pi^\#$, we have that the sequence $(J_{\pi^\#}(\pi_t))_t$ in non-increasing, and possesses a subsequence $(J_{\pi^\#}(\pi_{\tau'_t}))_t$ which converges to $J_{\pi^\#}(\pi^\#) = 0$, thus the whole sequence $(J_{\pi^\#}(\pi_t))_t$ converges to 0. This proves that π_t converges to the set of Nash equilibria in J_{π^*}-distance (as well as in exploitability using Lemma 2). □

M. Proof of Theorem 5

Theorem 5 Consider MAIO-BR with a ℓ_2-regularizer, and a learning rate $\eta_t = \frac{I(b_t, \pi_t)}{2c}$. Define

$$\varepsilon = \inf_{\pi^* \in \Pi^*, i \in \{1,2\}, x \in X_i, a \in A : \mu_i^{\pi^*}(x,a) > 0} \mu_i^{\pi^*}(x,a).$$

If $\varepsilon > 0$ then MAIO-BR converges to the set of Nash equilibria at an exponential rate.

Proof. Since $\varepsilon > 0$ we deduce from the proof of Lemma 5 that there exists constants $c' = 16/\varepsilon$ and $\delta = \frac{\varepsilon^2}{4|A|}$ such that as soon as $d(\pi^*, \pi_t) \leq \delta$, we have

$$J_{\pi^*}(\pi_t) \leq c' d(\pi^*, \pi_t), \tag{21}$$

for all Nash eq. $\pi^* \in \Pi^*$.

Now, from Theorem 4, π_t converges to Π^* in J_{π^*} distance. Let us write $\pi_t^* = \arg\min_{\pi^* \in \Pi^*} J_{\pi^*}(\pi_t)$. Thus for every player i, and every state $x \in X_i$ and action $a \in A$, we have that $\mu_i^{\pi_t^*}(x)[\pi_{i,t}^*(a|x) - \pi_{i,t}(a|x)]^2$ converges to 0. Now, for any $x \in X_i$, we have that $\mu_i^\pi(x, a)$ is the product of policy probabilities $\prod_{j=0}^n \pi_i(a_j|x_j)$ along the path $(x_0, a_0, x_1, a_1, \ldots, x_n = x, a_n = a)$, where $(x_j)_{0 \leq j \leq n} \in X_i$, thus

$$\begin{aligned}
\mu_i^{\pi_t^*}(x, a) - \mu_i^{\pi_t}(x, a) &= \prod_{j=0}^n \pi_{i,t}^*(a_j|x_j) - \prod_{j=0}^n \pi_{i,t}(a_j|x_j) \\
&= \sum_{j=0}^n \prod_{l=0}^{j-1} \pi_{i,t}^*(a_l|x_l)\big[\pi_{i,t}^*(a_j|x_j) - \pi_{i,t}(a_j|x_j)\big] \prod_{l=j+1}^n \pi_{i,t}(a_l|x_l) \\
&= \sum_{j=0}^n \mu_i^{\pi_t^*}(x_j)\big[\pi_{i,t}^*(a_j|x_j) - \pi_{i,t}(a_j|x_j)\big] \prod_{l=j+1}^n \pi_{i,t}(a_l|x_l).
\end{aligned}$$

Since each $\mu_i^{\pi_t^*}(x_j)[\pi_{i,t}^*(a_j|x_j) - \pi_{i,t}(a_j|x_j)]$ converges to 0, we deduce that $\mu_i^{\pi_t^*}(x, a) - \mu_i^{\pi_t}(x, a)$ converges to 0 as well. Thus π_t converges to Π^* in d-distance as well. Let us write $\tilde{\pi}_t^* = \arg\min_{\pi^* \in \Pi^*} d(\pi^*, \pi_t)$.

We deduce there exists a time T such that for all $t \geq T$, $d(\tilde{\pi}_t^*, \pi_t) \leq \delta$, thus from (21), we deduce that for $t \geq T$, $J_{\tilde{\pi}_t^*}(\pi_t) \leq c' d(\tilde{\pi}_t^*, \pi_t)$. Since from Lemma 4 we have $d(\tilde{\pi}_t^*, \pi_t) \leq \frac{1}{\kappa^2} I^*(\pi_t)^2$, thus

$$J_{\pi_t^*}(\pi_t) \leq J_{\tilde{\pi}_t^*}(\pi_t) \leq \frac{c'}{\kappa^2} I^*(\pi_t)^2.$$

Thus from Theorem 3, we deduce that

$$\begin{aligned}
J_{\pi_{t+1}^*}(\pi_{t+1}) &\leq J_{\pi_t^*}(\pi_{t+1}) \\
&\leq J_{\pi_t^*}(\pi_t) - \frac{1}{4c} I^*(\pi_t)^2 \\
&\leq J_{\pi_t^*}(\pi_t)\Big(1 - \frac{\kappa^2}{4cc'}\Big).
\end{aligned}$$

Thus π_t converges to the set of Nash equilibria at an exponential rate. □

N. Proof of Lemma 4

Lemma 4. There exists a constant $\kappa > 0$ (which depends on the game), such that for any policy π we have

$$I^*(\pi) \geq \kappa \min_{\pi^* \in \Pi^*} \sqrt{d(\pi^*, \pi)},$$

where

$$\begin{aligned}
d(\pi^*, \pi) &\overset{\text{def}}{=} \sum_i \|\mu^{\pi_i^*} - \mu^{\pi_i}\|^2 \\
&= \sum_i \sum_{x \in X_i, a \in A} \big[\mu_i^{\pi^*}(x, a) - \mu_i^\pi(x, a)\big]^2,
\end{aligned}$$

and $\mu_i^\pi(x, a) \overset{\text{def}}{=} \mu_i^\pi(x)\pi_i(a|x)$.

We first state a result which will be useful for the proof of Lemma 4.

Lemma 10 (Value function is multilinear)**.** *Under the perfect recall assumption, the value function V_i^π is linear w.r.t. the policy at each state. More precisely, for any $x \in X$, write $\pi_{x\to\pi'}$ the policy defined as being equal to π everywhere except in x where it is π', the mapping*

$$\pi' \mapsto V_i^{\pi_{x\to\pi'}}$$

is linear. Thus $\pi \mapsto V_i^\pi$ is a multilinear map.

Proof. Since at the history level the MDP has the structure of a tree (there is a unique path from the initial history to any history h), the value function V^π is a linear function of the policy $\pi(x)$ at each state x. Indeed we have

$$V_i^\pi = \sum_{h\in H}\sum_a \mu^\pi(h,a) r_i(h,a), \tag{22}$$

where $\mu^\pi(h,a) \stackrel{\text{def}}{=} \mu^\pi(h)\pi(a|x(h))$ is the probability of reaching the history-action (h,a). Now from the tree structure of the history MDP, the reach probability

$$\mu^\pi(h,a) = \prod_{j=0}^{n} p(h_j|h_{j-1},a_{j-1})\pi(a_j|x(h_j)),$$

(where we define $p(h_0|h_{-1},a_{-1}) \stackrel{\text{def}}{=} \rho_0(h_0)$) is the product of the transition probabilities $p(h_j|h_{j-1},a_{j-1})$ and policy $\pi(a_j|x_j)$ along the path $(h_j,a_j)_{0\le j\le n}$, where $h_n = h$ and $a_n = a$ for some $n \ge 0$ (n is the length of the path). Since the policy $\pi(x)$ appears at most once in this product, the probability $\mu^\pi(h,a)$ is a linear function of the policy at each state, thus the mapping $\pi' \mapsto \mu^{\pi_{x\to\pi'}}(h,a)$ is linear, and so is the value function V_i^π as written in (22). Thus the mapping

$$\pi' \mapsto V_i^{\pi_{x\to\pi'}}$$

is linear. □

Proof of Lemma 4. Our approach consists of extending the proof of Lemma 1, where now we consider player's individual reach probabilities μ_i instead of distribution over actions. We write $\Delta_i \subseteq \mathbb{R}^{X_i\times A}$ for the convex set of possible reach probabilities for each player $i = 1,2$: $\Delta_i = \{(\mu_i^\pi(x,a))_{(x,a)\in X_i\times A}$ for any possible $\pi\}$. Given a set of reach probabilities $\mu_i \in \Delta_i$, we also introduce the notation $\pi_i(\mu_i)$ for the behavioural policy that induces the reach probabilities μ_i. For any x such that $\sum_{b\in A}\mu_i(x,b) > 0$ we have that $\pi_i(\mu_i)(a|x) = \frac{\mu_i(x,a)}{\sum_{b\in A}\mu_i(x,b)}$. We then define, for any set of reach probabilities $(\mu_i \in \Delta_i)_{i\in\{1,2\}}$,

$$\begin{aligned}
f_1(\mu_1) &\stackrel{\text{def}}{=} \max_{\pi_1'}\min_{\pi_2'} V^{(\pi_1',\pi_2')} - \min_{\pi_2'} V^{(\pi_1(\mu_1),\pi_2')} \\
&= \max_{\pi_1'}\min_{\pi_2'} V^{(\pi_1',\pi_2')} - \min_{\pi_2'} J_1^{\pi_2'}(\mu_1), \\
f_2(\mu_2) &\stackrel{\text{def}}{=} \max_{\pi_1'} V^{(\pi_1',\pi_2(\mu_2))} - \min_{\pi_2'}\max_{\pi_1'} V^{(\pi_1',\pi_2')} \\
&= \max_{\pi_1'} J_2^{\pi_1'}(\mu_2) - \min_{\pi_2'}\max_{\pi_1'} V^{(\pi_1',\pi_2')},
\end{aligned}$$

where we have introduced the linear maps: $\mu_i \in \Delta_i \mapsto J_i^{\pi'_{-i}}(\mu_i)$ defined as

$$J_i^{\pi'_{-i}}(\mu_i) \stackrel{\text{def}}{=} \sum_{x\in X_i, a\in A} \mu_i(x,a)\mu_{\neq i}^{\pi'}(x) r_i^{\pi'_{-i}}(x,a) = V_i^{(\pi_i(\mu_i),\pi'_{-i})},$$

where $r_i^{\pi'_{-i}}(x,a)$ are the rewards of the MDP $\mathcal{M}_i^{\pi_{-i}}$ (see Appendix I). From the minimax theorem we have that $\max_{\pi_1'}\min_{\pi_2'} V^{(\pi_1',\pi_2')} = \min_{\pi_2'}\max_{\pi_1'} V^{(\pi_1',\pi_2')}$, thus $I^*((\pi_1(\mu_1),\pi_2(\mu_2))) = f_1(\mu_2) + f_2(\mu_1)$.

For any $i \in \{1,2\}$, we have that $f_i : \Delta_i \to \mathbb{R}$ is the game value, plus a maximum (over the set of policies π'_{-i}) of linear functions: $\mu_i \mapsto J_i^{\pi'_{-i}}(\mu_i)$, thus is a convex function. Now, from Lemma 10, for a given μ_i the mapping

$$\pi'_{-i} \mapsto J_i^{\pi'_{-i}}(\mu_i) = V_i^{(\pi_i(\mu_i),\pi'_{-i})}$$

is multilinear in π'_{-i}, thus the maximum of $J_i^{\pi'_{-i}}(\mu_i)$ over π'_{-i} is reached for (at least) a deterministic policy. Since the number of deterministic policies is finite (since the state and action spaces are finite), we have that f_i is the maximum –over a *finite number* of policies– of linear functions.

Thus f_i is a piecewise affine function defined on a partitioning of Δ_i (which is a convex polytope of $\mathbb{R}^{X_i \times A}$) where each piece of this partition is an intersection of half spaces, thus a convex polytope. A convex polytope also has the property of being defined as the convex hull of a finite number of extreme points (called vertices). Let M_i denote the (finite) set of reach probabilities μ_i which form the vertices of this polytope partitioning.

f_i is non-negative, so let us denote M_i^* the set of reach probabilities μ_i such that $f_i(\mu_i) = 0$. The corresponding policies $\pi(\mu_i)$ for $\mu_i \in M_i^*$ corresponds to the set of Nash equilibria for player i. M_i^* is also a (closed) convex polytope of Δ_i.

For any $\mu_i \in \Delta_i$, write $P_{M_i^*}(\mu_i)$ the ℓ_2-projection of μ_i onto M_i^*:

$$\begin{aligned} P_{M_i^*}(\mu_i) &\overset{\text{def}}{=} \operatorname*{arg\,min}_{\mu_i^* \in M_i^*} \|\mu_i - \mu_i^*\| \\ \text{where } \|\mu_i - \mu_i^*\|^2 &= \sum_{x \in X_i, a \in A} \big[\mu_i(x,a) - \mu_i^*(x,a)\big]^2. \end{aligned}$$

We define

$$\kappa_i = \min_{\mu_i \in M_i \setminus M_i^*} \frac{f_i(\mu_i)}{\|\mu_i - P_{M_i^*}(\mu_i)\|}.$$

Since M_i is a finite set, the minimum is reached for some value $\kappa_i > 0$.

Now consider any $\mu_i \in \Delta_i \setminus M_i^*$ and its projection $P_{M_i^*}(\mu_i)$. μ_i belongs to some convex polytope with corresponding vertices $\{\mu_i^j\}_j \in M_i$, thus $\mu_i = \sum_j \lambda_j \mu_i^j$, for some non-negative coefficients λ_i which sum to 1 (the so-called barycentric coordinates). Let us introduce $\mu'_i = \sum_j \lambda_j P_{M_i^*}(\mu_i^j)$ the barycenter of the projections of those vertices. Since M_i^* is convex, then $\mu'_i \in M_i^*$.

We have

$$\begin{aligned} \frac{f_1(\mu_i)}{\|\mu_i - P_{M_i^*}(\mu_i)\|} &\overset{(a)}{\geq} \frac{f_1(\mu_i)}{\|\mu_i - \mu'_i\|} \\ &\overset{(b)}{=} \frac{\sum_j \lambda_j f_i(\mu_i^j)}{\big\|\sum_j \lambda_j \big(\mu_i^j - P_{M_i^*}(\mu_i^j)\big)\big\|} \\ &\overset{(c)}{\geq} \frac{\sum_j \lambda_j f_i(\mu_i^j)}{\sum_j \lambda_j \big\|\mu_i^j - P_{M_i^*}(\mu_i^j)\big\|} \\ &\overset{(d)}{\geq} \kappa_i, \end{aligned}$$

where (a) results from the fact that $\|\mu_i - P_{M_i^*}(\mu_i)\| \leq \|\mu_i - \mu'_i\|$ by definition of the projection $P_{M_i^*}$, (b) from the linearity of f_i in the polytope containing μ_i, (c) from the triangle inequality, (d) from the definition of κ_i thus that for each j, $\frac{f_i(\mu_i^j)}{\big\|\mu_i^j - P_{M_i^*}(\mu_i^j)\big\|} \geq \kappa_i$. We deduce that

$$\begin{aligned} f_i(\mu_i) &\geq \kappa_i \|\mu_i - P_{M_i^*}(\mu_i)\| \\ &= \kappa_i \min_{\mu_i^* \in M_i^*} \sqrt{\sum_{x \in X_i, a \in A} \|\mu_i(x,a) - \mu_i^*(x,a)\|^2} \\ &= \kappa_i \min_{\pi_i^* \in \Pi_i^*} \sqrt{\sum_{x \in X_i, a \in A} \|\mu^{\pi_i(\mu_i)}(x,a) - \mu^{\pi_i^*}(x,a)\|^2}. \end{aligned}$$

We deduce that

$$\begin{aligned} I^*(\pi) &= \sum_{i\in\{1,2\}} f_i(\mu_i(\pi_i)) \\ &\geq \sum_{i\in\{1,2\}} \kappa_i \min_{\pi_i^*\in\Pi_i^*} \sqrt{\sum_{x\in X_i, a\in A} \|\mu_i^{\pi}(x,a) - \mu^{\pi_i^*}(x,a)\|^2} \\ &\geq \kappa \min_{\pi^*\in\Pi^*} \sqrt{\frac{1}{2}\sum_{i\in\{1,2\}}\sum_{x\in X_i, a\in A} \|\mu^{\pi_i(\mu_i)}(x,a) - \mu^{\pi_i^*}(x,a)\|^2} \\ &= \kappa \min_{\pi^*\in\Pi^*} \sqrt{d(\pi^*,\pi)}, \end{aligned}$$

for $\kappa = \sqrt{2}\min(\kappa_1,\kappa_2)$. □

O. Conjecture about MAIO-BR with entropy regularizer

We conjecture that MAIO-BR with entropy regularizer can achieve an exponential convergence if and only if (at least) one Nash eq. is an interior point (i.e. a strictly stochastic policy).

In order to give some argument to this claim, let us first derive a 'reverse' Pinsker's inequality. For any $p, q \in \Delta(A)$, we have

$$\begin{aligned} KL(p,q) &= \sum_a p(a)\log\frac{p(a)}{q(a)} \\ &= \sum_a p(a)\log\big(1+\frac{p(a)-q(a)}{q(a)}\big) \\ &\leq \sum_a \frac{p(a)}{q(a)}\big(p(a)-q(a)\big) \\ &= \sum_a \frac{\big(p(a)-q(a)\big)^2}{q(a)} \\ &\leq \frac{\sum_a \big(p(a)-q(a)\big)^2}{\min_a q(a)}. \end{aligned}$$

From this inequality we deduce that

$$KL(\pi^*,\pi) \leq \frac{\|\pi-\pi^*\|_2^2}{\min_a \pi(a)}.$$

Thus assuming all Nash eq. π^* are interior points, we could prove that all π_t are bounded away from the boundary of the simplex (i.e., $\min_{i\in\{1,2\}}\min_{a\in A_i}\min_t \pi_{i,t}(a) \geq c$ for some constant $c>0$ which depends on the initial policy π_0 only) thus Lemma 1 would imply that

$$I^*(\pi_t) \geq \kappa\|\pi^*-\pi_t\|^2 \geq \Big(\min_{i\in\{1,2\}}\min_{a\in A_i}\pi_{i,t}(a)\Big)KL(\pi^*,\pi_t) \geq \kappa\, cKL(\pi^*,\pi_t),$$

which plugged into Theorem 1 will imply exponential convergence of MAIO-BR with entropy regularization.

However in the case when all Nash equilibria are on the boundary of the simplex, this property may not hold and we conjecture the algorithm will suffer from a lower convergence rate. We leave this open question for future work.

P. More experiments

P.1. MAIO-BR-IIG versus CFR, CFR+, CFR-BR

We test MAIO-BR-IIG on four imperfect information games from the OpenSpiel library (Lanctot et al., 2019):

- **Kuhn poker** is a simplified poker game with only 3 cards, first proposed in (Kuhn, 1950), with only 12 information states.
- **Leduc poker** consists of two rounds of bets with a 6-card deck in two suits, e.g. JS,QS,KS, JH,QH,KH. It contains 936 information states.
- **Liars Dice(1,1)** is dice game where each player gets a single private die, rolled at the start of the game, and players proceed to bid on the outcomes of all dice in the game. It contains 24576 information states.
- **Goofspiel** is a card game where players obtain point cards by bidding simultaneously. We use an imperfect information variant containing 162 information states where bid cards are not revealed.

We compare MAIO-BR against the following algorithms:

- **CFR** (Counterfactual Regret Minimization) (Zinkevich et al., 2008): the uniform mixture of all past policies (the so-called average policy) converges in value to the value of the Nash eq. with a rate $O(1/\sqrt{t})$.
- **CFR-BR** (Johanson et al., 2012) performs CFR updates using a best-response agent as its opponent. In addition to having the average policies converging, the last policy also converges (in value) to the value of the Nash eq. with high probability, also with a $O(1/\sqrt{t})$ rate. We reported CFR-BR with average policy since CFR-BR with the last policy did not perform better (and was less stable) than the average policy in our runs.
- **CFR+** (Tammelin et al., 2015) combines 3 improvements over CFR: (a) linear-time average of the strategy (i.e. the most recent strategies are weighted more), (b) alternating-updates, (c) use regret-matching+ instead of regret-matching for the basic regret minimization algorithm. Although CFR+ has a theoretical guarantee of a $O(1/\sqrt{t})$ convergence rate as well, it numerically performs much better than this rate and is usually considered the state-of-the-art method.

Figure 2 reports the exploitability $I^*(\pi_t) \max_\pi I(\pi, \pi_t)$ for MAIO-BR using ℓ_2 and entropy regularization compared to CFR, CFR-BR, and CFR+. We notice MAIO-BR-ℓ_2 greatly outperforms the other methods in Kuhn Poker and Goofspiel. However CFR+ works better than MAIO-BR on Liar's Dice and Leduc Poker. We believe the combination of the 3 ingredients contained in CFR+ contribute to a better rate than theoretically reported ($1/\sqrt{t}$ rate) for CFR+. Also we do not observe the exponential rate for MAIO-BR within the 10^{-6} iterations of the algorithm. It could be that the constant κ characterizing the exponential rate of MAIO-BR is too small or there may be several Nash equilibria (in which case Theorem 5 would not apply).

The difference of the number of steps in Figure 2 between the different curves is due to both (a) the unequal run time given to each methods (the baseline ran for much longer) and (b) the less efficient implementation for MAIO, which performs several tree-traversal per update.

Experiment reproducibility: For CFR, CFR+ and CFR-BR, the exploitability is computed every 100 steps, while for the MAIO curves, there is one data point per step up to 1000 iterations and then one data-point every 100 steps, which explains the change in the curve smoothness. There is no window averaging or any post-processing of the raw data.

In all experiments we used the ℓ_2-projection (Chen and Ye, 2011) for MAIO with the ℓ_2-regularizer, available at `https://github.com/swyoon/projsplx`. We had to replace `eq_idx = np.searchsorted(t_iter - y_iter, 0, side="left")` with `eq_idx = ((t_iter - y_iter) >= 0).nonzero()[0][0]` because the first one was giving incorrect results.

P.2. Best response versus greedy policy

In Figure 3 we report the exploitability $\max_{\pi'} I(\pi', \pi_t)$ (left column) and the improvement $I(\bar{\pi}_t, \pi_t)$ (right column) for two choices of improved policies $\bar{\pi}_t$: the best response b_t and the greedy policy g_t. Note that for the best-response, the improvement is the same as the exploitability.

For Kuhn Poker, the best-response and greedy policy are most of times the same (except around step 1159, hardly-visible on the graphs) because of the very small and shallow size of the tree (there is a single action taken by the second player, only the first player can plays 2 times when the first player passes and the second bets).

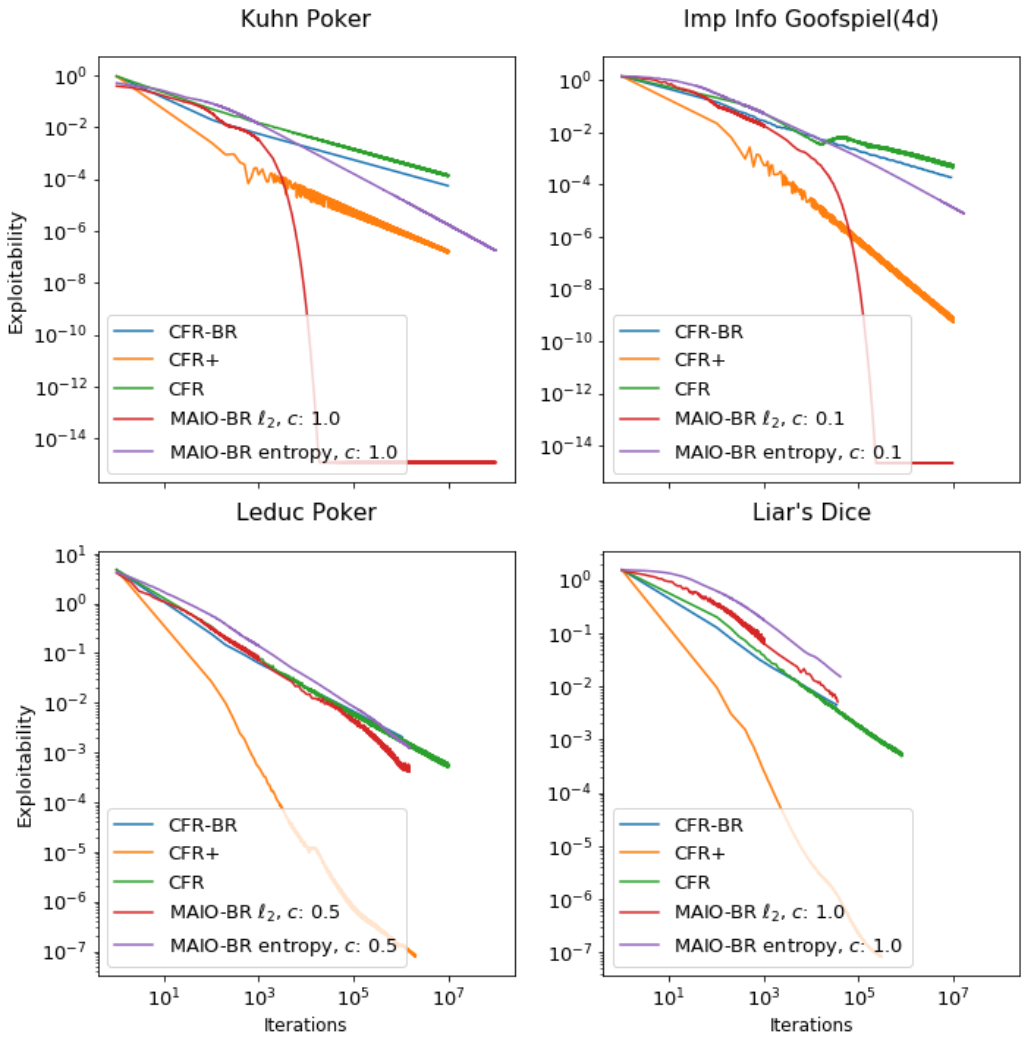

Figure 2. MAIO-BR with ℓ_2 and entropy regularization in four IIGs compared to CFR, CFR-BR with current policy, CFR-BR with average policy, and CFR+.

For the greedy policy, computing a single step of greedy policy does not necessarily produces a strictly positive improvement. Indeed it could be that $I(g_t, \pi_t) = 0$ while not being at a Nash eq. yet. The reason is that although the greedy policy locally improves the current policy and generates a local improvement $I(g_t, \pi_t)(x) > 0$ at some $x \in X_i$, if the state x has no probability to be reached under the policy $(g_{i,t}, \pi_{-i,t})$ (i.e., $\mu^{(g_{i,t}, \pi_{-i,t})}(x) = 0$), then this local improvement does not impact the global improvement $I(g_t, \pi_t)$, see Lemma 6. So if we select our learning rate proportional to the global improvement and this one is 0 then no learning occurs. This is actually what happens for Goofspiel (see the plateau in Fig. 3, bottom left curve).

Thus, the plateau observed in Goofspiel is the result of (a) the fact that we are using ℓ_2 regularization which makes it possible to have unreachable states (this is not the case when using the entropy regularizer as in that case the policy π_t assigns strictly positive mass to all actions), (b) the fact that the greedy policy may lead to zero improvement while not being at the Nash, (c) the use of a learning rate $\eta_t = cI(g_t, \pi_t)$ proportional to the improvement.

To circumvent this problem, we can either choose a learning rate independent of the improvement, or use the entropy regularizer instead of ℓ_2. We also designed the following procedure to confirm our analysis: we iteratively compute a sequence of greedy policies until the corresponding improvement is strictly positive (which will always happen in at most n iterations, where n is the length of longest history). This version is called MAIO-Iterative-Greedy. These alternative solutions are presented in Figure 4.

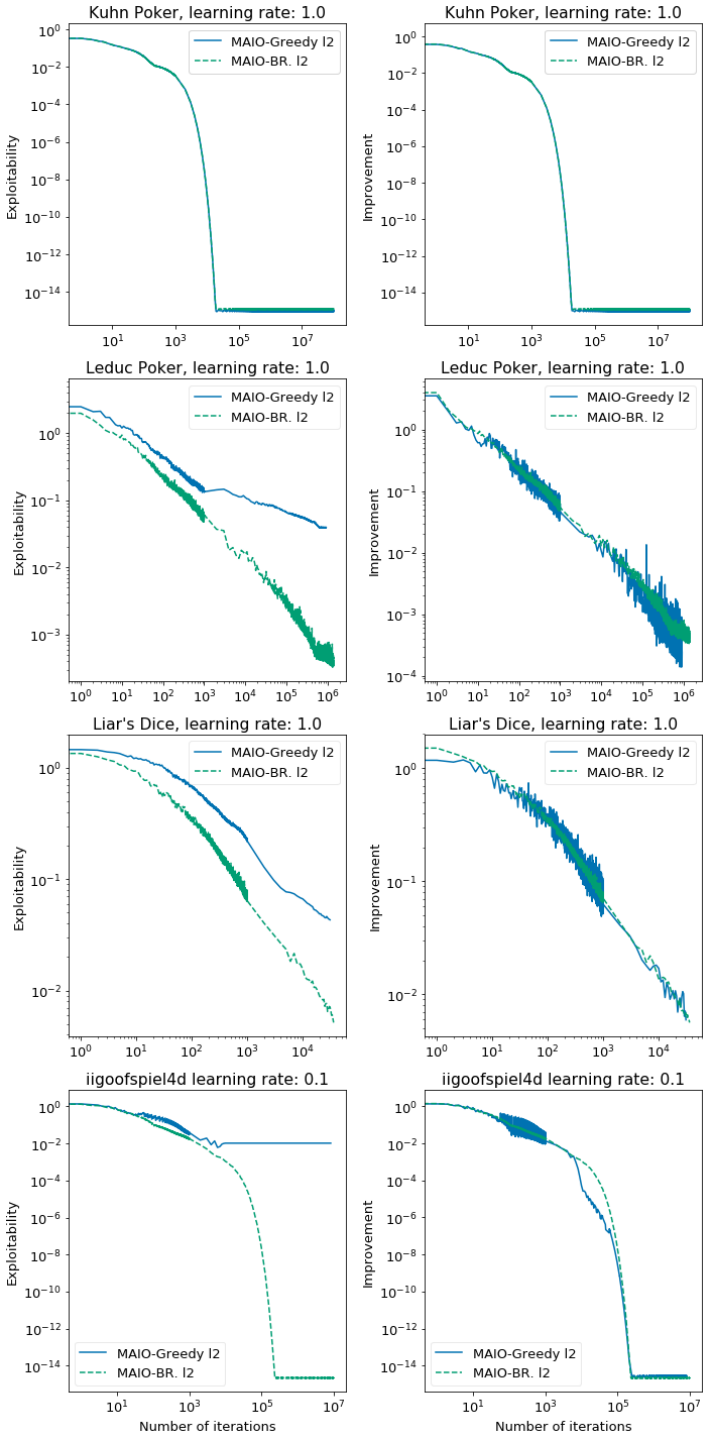

Figure 3. Exploitability $I^*(\pi_t)$ (left column) and improvement $I(\bar{\pi}_t, \pi_t)$ (right column) of the improved policy $\bar{\pi}_t$ for the best-response $b_t \in \arg\max_\pi I(\pi, \pi_t)$ and greedy policy $g_t(x) = \arg\max_a Q^{\pi_t}(x, a)$. The plateau in Goofspiel is explained in the main text.

P.3. Mixture best-response

Here we report the exploitability of MAIO when we use as improved policy a mixture between the best response and the current policy (as defined in (10)). The reason we report these experiments is to illustrate the impact on the convergence speed of using as improved policy a policy which is in between an improved policy (here the best response) and the current

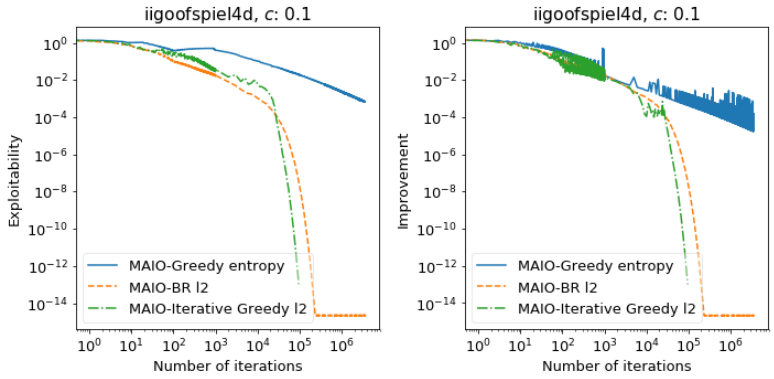

Figure 4. Exploitability $I^*(\pi_t)$ (left column) and improvement $I(\bar{\pi}_t, \pi_t)$ (right column) when using the Greedy improvement with entropy regularization or when using the Greedy iterative improvement with ℓ_2 regularization.

policy. The benefit of using a such mixture would be to use off-policy policy evaluation methods to evaluate $Q^{(\pi_{i,t}, \bar{\pi}_{-i,t})}$ the value function of each player against the improved opponent, while generating games according to the current policy π_t. We observe in Figure 5 that, as expected, the speed at which the exploitability goes to zero depends on the coefficient α of the mixture. We notice that when $\alpha = 0$ (we use for $\bar{\pi}_t$ the current policy π_t) the algorithm does not converge, since there is no improvement. It is interesting to notice that as soon as $\alpha > 0$ then convergence occurs, even for very small values of α (for example for $\alpha = 0.1$).

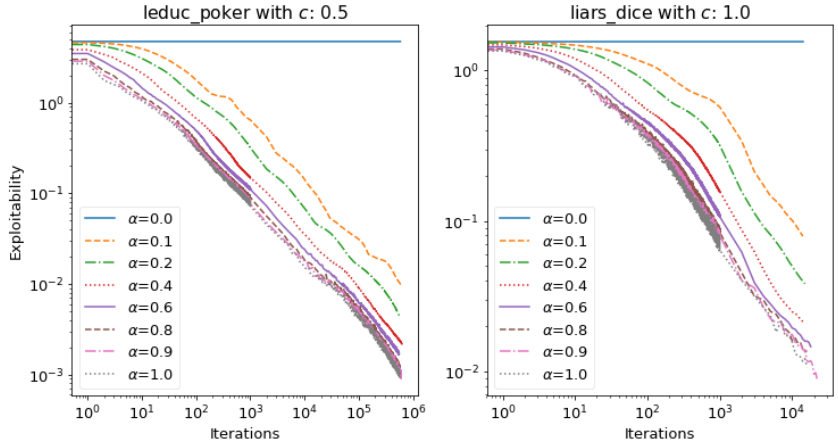

Figure 5. Exploitability $I^*(\pi_t)$ of MAIO-BR-mixture using as the improved policy $\bar{\pi}_t = (1 - \alpha)b_t + \alpha\pi_t$ a mixture between the current policy π_t and the best-response b_t.