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+).

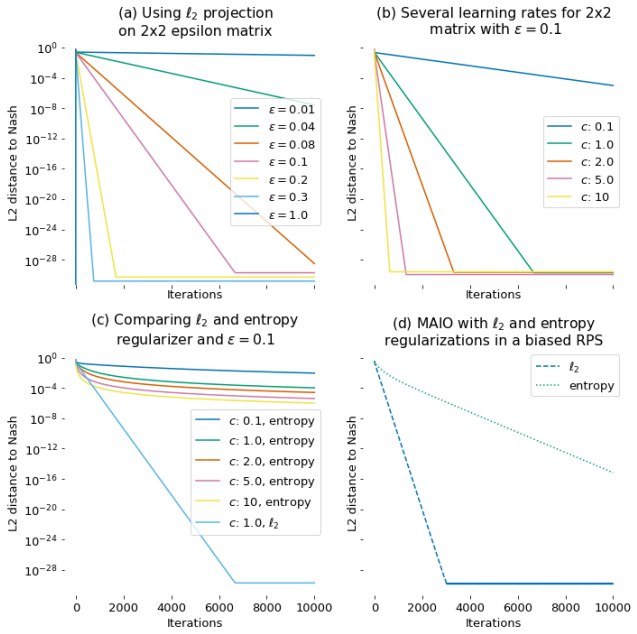

Figure 1. We report ℓ_2 distance to the Nash eq. (in log-scale) for MAIO-BR on the ε-matrix game (Fig. a,b,c) and the biased rock-paper-scissors game (Fig. d). MAIO-BR with ℓ_2 regularization shows an exponential convergence whose rate depend on ε (Fig. a) and the constant c (Fig. b) used in the learning 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)*.