Further Developments of Extensive-Form Replicator Dynamics using the Sequence-Form Representation

Marc Lanctot
Department of Knowledge Engineering, Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands
marc.lanctot@maastrichtuniversity.nl

ABSTRACT

Replicator dynamics model the interactions and change in large populations of agents. Standard replicator dynamics, however, only allow single action interactions between agents. Complex interactions, as modeled by general extensive games, have received comparatively little attention in this setting. Recently, replicator dynamics have been adapted to extensive-form games represented in sequence form, leading to a large reduction in computational resource requirements. In this paper, we first show that sequence-form constraints and realization equivalence to standard replicator dynamics are maintained in general n-player games. We show that sequence-form replicator dynamics can minimize regret, leading to equilibrium convergence guarantees in two-player zero-sum games. We provide the first empirical evaluation of sequence-form replicator dynamics, applied to n-player Kuhn poker with two, three, and four players. Our results show that the average strategies generated by sequence-form replicator dynamics produce approximate equilibrium strategies with increasing accuracy over time.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

General Terms

Algorithms, Economics, Theory

Keywords

Replicator dynamics; game theory; extensive-form games; multiplayer; Nash equilibrium; sequence form

1. INTRODUCTION

Evolutionary game theory [17, 8] has been used to explain complex interactions in multiagent systems such as population dynamics [11], animal behavior [18], and multiagent learning [25]. The most popular and widely-studied population dynamic is the so-called *replicator dynamic* [24]. Replicator dynamics quantify increases in the proportion of individuals in a population based on their relative fitness levels as determined through payoffs of competitive games. Agents with higher fitness replicate more than agents with lower fitness and the resulting process leads to an evolutionary game that models population change.

Multiagent systems that evolve under replicator dynamics have desirable properties and connections to classical game theory. For example, a fixed point of a population under replicator dynamics corresponds to a Nash equilibrium of the underlying game and stability of the system can be analyzed using theory of dynamical systems [5, 8]. In addition, average payoff of a population increases [11], dominated strategies do not survive [8], and stable equilibrium points correspond to trembling-hand perfect equilibria [5, 22].

In the classic setup, the underlying stage game is a symmetric normal-form game. This limits the interaction among the agents since payoffs are determined from single decisions made by each agent. In general, the games played among agents can be more complex, such as multi-step sequences of actions as in extensive-form games. When faced with this added complexity, one option is to convert the extensive-form game to its equivalent normal-form, but this is only possible for small games. Another option is to heuristically abstract the strategy space [20], but abstraction is lossy and may lead to loss of desired theoretical properties. Evolutionary dynamics have been extended to multistage models such as extensive games [3] and stochastic games [4, 9], however the focus has been mainly on subgame-decomposable formalisms such as perfect-information games and simultaneous-move games. In the general setting of imperfect information games, the extensive-form game may not necessarily be decomposed into smaller subgames.

Recently, efficient replicator dynamics have been proposed for the general case of extensive-form games with imperfect information [6], based on sequence-form representations [13]. In their paper, the authors introduce discrete and continuous time sequence-form replicator dynamics (SFRD) that can represent general extensive games with much less computational requirements than their normal form.

In this paper, we first show that important properties are conserved when SFRD are employed in games with more than two players. We show that SFRD leads to a specific form of no-regret learning and hence convergence to an equilibrium can be guaranteed in two-player zero-sum games. We present the first empirical evaluation of SFRD and, in particular, the first evidence confirming the convergence of

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

SFRD to a Nash equilibrium, both in theory (for two player zero-sum) and practice (for n players). We also show empirical evidence of average strategies converging to equilibrium in all cases, despite persistent changes in the approachability of the strategies modified by the evolutionary dynamics.

2. GAME THEORY BACKGROUND

In this section we define the relevant game-theoretic terminology that forms the basis of our analysis. The notation used here is based on [19]. For a comprehensive introduction and survey of the fundamental topics, see [22].

An extensive-form game models sequential decision making. There are n decision-making agents called **players** $i \in N = \{1, \ldots, n\}$. In turn, players choose **actions** leading to sequences called **histories** $h \in H$. A history $z \in Z$, where $Z \subseteq H$, is called a **terminal history** and represents a fully-played game from start to finish. At each terminal history z there is a payoff $u_i(z)$ in $[0, 1]$ to each player i. At each nonterminal history h, there is a single player to act, $P : H \backslash Z \to N \cup \{c\}$ where c is a special player called **chance** (also sometimes called nature) that plays with a fixed stochastic strategy; *e.g.* chance is used to represent dice rolls and card draws. The game starts in the empty history, and at each step, given the current history h, $P(h)$ chooses an action $a \in A(h)$ leading to successor history $h' = ha$; in this case we call h a **prefix** of h' and denote this relationship by $h \sqsubset h'$. Also, for all $h, h', h'' \in H$, if $h \sqsubset h'$ and $h' \sqsubset h''$ then $h \sqsubset h''$. Each set N, H, Z, and $A(h)$ is finite and every history has finite length.

Define $\mathcal{I} = \{\mathcal{I}_i \mid i \in N\}$ the set of information partitions. $\mathcal{I}_i$ is a partition over $H_i = \{h \mid P(h) = i\}$ where each part is call an **information set**. Intuitively, an information set $I \in \mathcal{I}_i$ that belongs to player i represents a state of the game with respect to what player i knows. Each I is a set of histories that a player cannot tell apart due information hidden from that player. For all $h, h' \in I$, $A(h) = A(h')$ and $P(h) = P(h')$; hence, often we use $A(I)$ and $P(I)$.

We also define the **choice set** of (information set, action) pairs for one player to be $Q_i = \{(I, a) \mid I \in \mathcal{I}_i, a \in A(I)\} \cup \{q_\emptyset\}$, where $q_\emptyset$ is the empty(root) choice. For a history $h \in H$, define $X_i(h) = (I, a), (I', a'), \cdots$ to be the sequence of player i's (information set, action) pairs (choices) that were encountered and taken to reach h in the same order as they are encountered and taken along h. In this paper, every extensive-form game has **perfect recall**, which means $\forall i \in N, \forall I \in \mathcal{I}_i : h, h' \in I \Rightarrow X_i(h) = X_i(h')$. Intuitively, this means that player i does not forget any information that they discovered during their play up to h. Denote $succ_i(I, a)$ the set of successor choices of player i, that is all (I', a') such that $X_i(h') = X_i(h),\ (I', a')$ where $h \in I, h' \in I'$.

A **behavioral strategy** for player i is a function mapping each information set $I \in \mathcal{I}_i$ to a probability distribution over the actions $A(I)$, denoted $\sigma_i(I)$. If every distribution in the range of this mapping assigns all of its weight on a single action, then the strategy is called **pure**. A **mixed** strategy is a single explicit distribution over pure strategies. Given a profile σ, we denote the probability of reaching a terminal history z under σ as $\pi^\sigma(z) = \prod_{i \in N} \pi_i(z)$, where each $\pi_i(z)$ is a product of probabilities of the actions taken by player i in $X_i(z)$. We use $\pi_i^\sigma(h, z)$ and $\pi^\sigma(h, z)$ to refer to the product of only those probabilities along the sequence from h to z, where $h \sqsubset z$. Define Σ_i to be the set of behavioral strategies for player i. As is convention, σ_{-i} and π_{-i}^σ refer to player $i's$ opponents' strategies and products (including chance's). An **ϵ-Nash equilibrium**, σ, is a set of $\sigma_i, \forall i \in N$ such that the benefit to switching to some alternative σ_i',

$$\max_{\sigma_i' \in \Sigma_i} \left\{ \sum_{z \in Z} \pi_i^{\sigma'}(z) \pi_{-i}^{\sigma}(z) u_i(z) \right\} - u_i(\sigma) \le \epsilon \qquad (1)$$

holds for each player $i \in N$. When $\epsilon = 0$, the profile is simply called a Nash equilibrium. In this paper, we assume payoffs $0 \le u_i(z) \le 1$. Payoffs in games outside this range can be shifted by a constant and then scaled by the payoff range without changing the set of strategies that optimize Equation 1. When $|N| = 2$ and $u_1(z) + u_2(z) = k$ for all $z \in Z$, then the game is a two-player k-sum game, where k is a constant; these games form an important subset of extensive-form games due to their worst-case guarantees: different equilibrium strategies result in the same expected payoff against any arbitrary opponent equilibrium strategy.

2.1 Sequence-Form Replicator Dynamics

The sequence-form was introduced by Koller, Megiddo and von Stengel as an efficient way to construct linear programs and complementarity problems for solving extensive-form games with perfect recall [13]. Rather than using a game's equivalent normal-form representation, the sequence-form imposes constraints compactly by using the game tree's structure, resulting in an exponentially smaller optimization problem. Define a **realization plan**, denoted $\mathbf{x}_i$, as a mapping from each $q \in Q_i$ to a **realization weight** $x_i(q) \in [0, 1]$ under the constraints that each nonterminal $x_i(q) = \sum_{q' \in succ_i(q)} x_i(q')$ and root weight $x_i(q_\emptyset) = 1$. Every realization plan has an equivalent behavioral strategy due to perfect recall.

Sequence-form replicator dynamics (SFRD) were recently introduced by Gatti, Panozzo, and Restelli [6]. Denote the realization profile $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. In the common special case of two players, each realization plan is represented as a vector and payoffs for each game outcome as a sparse payoff matrix $\mathbf{U}_i$, and so the expected utility is simply $u_i(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{U}_i \mathbf{x}_2$. In general, the expected utility to player i is

$$u_i(\mathbf{x}) = \sum_{q_1 \in Q_1, \cdots, q_n \in Q_n} \prod_{k=1}^{n} x_k(q_k) u_i(q_1, \ldots, q_n), \qquad (2)$$

where $u_i(q_1, \ldots, q_n) = 0$ if the combined choices are inconsistent with each player's public information or do not lead to a terminal history, otherwise equals the utility to player i given these choices multiplied by the probability of chance realizing the outcomes consistent with these choices.

Discrete-time SFRD starts with an arbitrary strategy $\mathbf{x}_i$ and, at each time step t, for all players $i \in N$ and all choices $q \in Q_i$, updates the weights using

$$x_i(q, t+1) = x_i(q, t) \frac{u_i(\mathbf{x}_{i \to g_q})}{u_i(\mathbf{x})}, \qquad (3)$$

where $\mathbf{x}_{i \to g_q}$ corresponds to the realization profile $\mathbf{x}$ except player i uses $g_q(\mathbf{x}_i)$ instead of $\mathbf{x}_i$. Here, $g_q(\mathbf{x}_i)$ returns a transformed realization plan that is explained below. Continuous-time SFRD is described by the differential equation for all players $i \in N$ and all $q \in Q_i$:

$$\dot{x}_i(q, t) = x_i(q, t) u_i(\mathbf{x}_{i \to \Delta g_q}), \qquad (4)$$

where $\mathbf{x}_{i \to \Delta g_q}$ corresponds to the profile $\mathbf{x}$ except player i uses $\Delta g_q(\mathbf{x}_i) = g_q(\mathbf{x}_i) - \mathbf{x}_i$ in the update equation.

The function $g_q(\mathbf{x}_i)$ modifies $\mathbf{x}_i$ in the following way. For for all choices $(I, a) \in X_i(q)$: the action a is always taken (realization weight set to 1) and actions $b \in A(I), b \neq a$ never taken (realization weight set to 0 and all child weights of (I, b) also set to 0). Every other $q' \in Q_i$ that does not directly contradict actions taken in $X_i(q)$ (*i.e.* due to actions taken by opponents or being a longer path than $X_i(q)$), $x_i(q')$, is renormalized. In essence, $g_q(\mathbf{x}_i)$ is a projection of $\mathbf{x}_i$ to its purest form based on q while still retaining the sequence-form constraints. Let $q = (I, a)$, one can equivalently think of $g_q(\mathbf{x}_i)$ as the realization plan where player i plays to make choice q (reach I and take action a), and otherwise plays $\mathbf{x}_i$. Specifically, for a given element q',

$$g_q(\mathbf{x}_i, q') = \begin{cases} 1 & \text{if } q' \in X_i(q), \\ \frac{x_i(q')}{\text{Ancestor}(q,q')} & \text{if } X_i(q) \sqsubseteq X_i(q'), \\ 0 & \text{otherwise}, \end{cases} \qquad (5)$$

where $\text{Ancestor}(q, q') = x_i(q'')$ and q'' is the last choice in the overlapping subsequence $X_i(q) \cap X_i(q')$. For some examples of how g_q changes $\mathbf{x}_i$, see [6].

An important result is that SFRD is realization-equivalent to the standard replicator dynamics. Therefore, applying SFRD is identical to applying standard replicator dynamics to the normal-form equivalent game. However, SFRD requires exponentially less space to represent the game.

2.2 Regret Minimization

Suppose an algorithm A must choose one action among K possible actions. Associated with each action is a payoff, generated from a distribution or by an adversary. The (possibly randomized) algorithm repeats the process over T trials and collects payoff u_k^t for choosing action k at time t. The (expected) **regret** is difference between the cumulative payoff and the payoff that would have been achieved by choosing the best single action in hindsight,

$$R^T = \mathbb{E}\left[\max_{k \in \{1,\ldots,K\}} \sum_{1 \le t \le T} u_k^t - \sum_{1 \le t \le T} u_{A(t)}^t\right] \qquad (6)$$

Define the average regret to be $\bar{R}^T = R^T/T$. We call algorithm A a **regret minimization algorithm** if $\lim_{T\to\infty} \bar{R}^T = 0$. Often, the algorithm modifies how it chooses during the trials based on the collected payoffs, and so these are also appropriately known as *no-regret learning algorithms.*

The Polynomial Weights (PW) algorithm [2] is a generalization of the Randomized Weighted Majority algorithm [16]. Each action has a weight w_k, initially set to 1. PW chooses action k with a probability $p_k = w_k / \sum_{k'=1}^{K} w_{k'}$. After each trial, the weights are updated using $w_k \leftarrow w_k(1 - \eta l_k))$, where l_k is a loss incurred from not choosing k on the last step and η is a learning rate parameter. Here, l_{min} is the loss of the single best action for a given loss sequence, i.e. the argument optimizing Equation 6. When $\eta \le \frac{1}{2}$, the regret of PW at time T is bounded by $R^T \le \eta Q_{min}^T + (\ln K)/\eta$, where $Q_{min}^T = \sum_{t=1}^{T} (l_{min}^t)^2$.

There are important connections between game theory and regret minimization [1]. One main result is that in two-player zero-sum games, if after T trials each average regret $\bar{R}_i^T \le \epsilon$, then the average profile $\bar{\sigma}^T$ corresponds to a 2ϵ-equilibrium.

Counterfactual Regret (CFR) is a notion of regret at the information set level for extensive-form games [26]. Suppose player i plays with strategy σ_i. The **counterfactual value** of taking action $a \in A(I)$ at information set I is the expected payoff when I is reached given that player i played to reach I and the opponents played σ_{-i},

$$v_i(I, \sigma, a) = \sum_{(h,z) \in Z_I} \pi_{-i}^{\sigma}(z) \pi_i^{\sigma_{I \to a}}(h, z) u_i(z), \qquad (7)$$

where $Z_I = \{(h, z) | z \in Z, h \in I, h \sqsubset z\}$, and $\sigma_{I\to a}$ is identical to σ except at I action a is taken with probability 1. The CFR algorithm places a regret minimizer at each $I \in \mathcal{I}_i$ which treats $v_i(I, \sigma, a)$ as the payoff for taking action a. The main result is that the combination of individual regret minimizers also minimizes overall average regret, and hence $\bar{\sigma}^T$ is a 2ϵ-equilibrium, with $\epsilon \to 0$ as $T \to \infty$.

3. FURTHER DEVELOPMENTS OF SFRD

In this section, we describe new developments and analyses of sequence-form replicator dynamics. First, we show that the generalization to $n > 2$ players preserves the theoretical properties proved in the original work. Second, we show that discrete-time SFRD can minimize a form of counterfactual regret leading to equilibrium convergence guarantees in two-player zero-sum games.

Note that these are two separate, mostly independent, developments. However, both of the following subsections provide a basis for the the empirical evaluation in Section 4.

3.1 More Than Two Players

In this subsection, we show that general n-player SFRD maintains sequence-form constraints and is realization equivalent to n-player normal-form replicator dynamics.

Overall, the analysis here is based on the previous one in [6] with some adjustments.

Definition 1. *A choice $q \in Q_i$ for player i is* ***reachable*** *under $\mathbf{x}_i$ if and only if $x_i(q) > 0$.*

We will restrict our analysis to reachable choices of $\mathbf{x}_i$. The behavior for unreachable parts of $\mathbf{x}_i$ is irrelevant since the expected payoff of any profile including it is unaffected over all possible opponent strategies.

Theorem 1. *Given a valid realization profile $\mathbf{x}(t) = (\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t))$, a new $\mathbf{x}(t + 1)$ produced by Equation 3 satisfies the sequence-form constraints.*

Proof. First, observe $x_i(q', t) = 0 \Rightarrow x_i(q', t+1) = 0$ for all reachable choices $q' \in Q_i$. Similarly, $g_{q_\emptyset}(\mathbf{x}_i, q') = \mathbf{x}_i(q')$ since $q_\emptyset$ is a subsequence of every choice $q \in Q_i$. So the root weight $x_i(q_\emptyset, t + 1) = x_i(q_\emptyset, t) = 1$. We will prove by induction that

$$x_i(q, t+1) = \sum_{q' \in succ_i(q)} x_i(q', t+1).$$

We assume that the constraints hold at time t. Applying Eq. 3 and multiplying each side by $u_i(\mathbf{x})$ gives

$$x_i(q, t) u_i(\mathbf{x}_{i \to g_q}, t) = \sum_{q' \in succ_i(q)} x_i(q', t) u_i(\mathbf{x}_{i \to g_{q'}}, t)$$

Recalling that the opponent strategies are fixed, we rewrite the utility $u_i(\mathbf{x}_{i\to g_q}, t)$ using Equation 2 as a dot product of components belonging to i and to the opponents $-i$:

$$\begin{aligned} u_i(\mathbf{x}_{i\to g_q}) &= g_q(\mathbf{x}_i, q_i')u_i(q_1',\ldots,q_n')\prod_{q_k',k\neq i} x_k(q_k') \\ &\quad + g_q(\mathbf{x}_i, q_i'')u_i(q_1'',\ldots,q_n'')\prod_{q_k'',k\neq i} x_k(q_k'') \\ &\quad \ldots \\ &= \mathbf{g}_q(\mathbf{x}_i)\cdot\mathbf{u}(\mathbf{x}_{-i}), \end{aligned}$$

and substituting from above we have

$$x_i(q,t)(\mathbf{g}_q(\mathbf{x}_i)\cdot\mathbf{u}(\mathbf{x}_{-i})) = \sum_{q'\in succ_i(q)} x_i(q',t)(\mathbf{g}_{q'}(\mathbf{x}_i)\cdot\mathbf{u}(\mathbf{x}_{-i})),$$

which, by distributivity and commutativity, rearranges to

$$\mathbf{u}(\mathbf{x}_{-i})\cdot(x_i(q,t)(\mathbf{g}_q(\mathbf{x}_i)) = \mathbf{u}(\mathbf{x}_{-i})\cdot\sum_{q'\in succ_i(q)} x_i(q',t)\mathbf{g}_{q'}(\mathbf{x}_i).$$

The vectors on the left side of the dot products are equal. Therefore, it suffices to show the vectors on the right side are also equal, which is similar to the case of two players. For choices q'' where $X_i(q'')$ is inconsistent with $X_i(q)$, the parent and child weights are all set to 0, so these elements are equal by the induction assumption. Similarly, for choices q'' where $q'' \in X_i(q)$, $g_q(\mathbf{x}_i, q'') = g_{q'}(\mathbf{x}_i, q'') = 1$, and these elements are also equal by the induction assumption. The remaining elements are equal by Eq. 5 and [6, Lemma 6]. □

We now discuss realization equivalence with n players.

Definition 2. *Two strategies σ_i and σ_i' are **realization equivalent** if $\forall h \in H, \forall \sigma_{-i} \in \Sigma_{-i}, \pi^{\sigma}(h) = \pi^{\sigma'}(h)$.*

In other words, every history is reachable with the same probability given an arbitrary fixed opponent profile. Due to perfect recall, every mixed strategy has an equivalent behavioral strategy [15]. Similarly, every mixed strategy has an equivalent realization plan and every realization plan has an equivalent behavioral strategy [13]. Therefore, the definition can be used for these other forms by reasoning about their equivalent behavioral forms.

An important result from [6] is that strategies produced by SFRD are realization equivalent to the standard normal-form replicator dynamics. In our n-player analysis, we reuse a key result ([6, Lemma 9]).

Theorem 2. *Given some game Γ with player set N, let $\rho(t)$ be a mixed strategy profile and $\rho(t+1)$ be the profile produced by standard discrete-time replicator dynamics using Γ's normal-form representation. Let $\mathbf{x}(t)$ be a realization profile and $\mathbf{x}(t+1)$ a realization profile in Γ's sequence-form produced by Equation 3. Then for all players $i \in N$, $\mathbf{x}(t+1)$ and $\rho(t+1)$ are realization equivalent.*

Proof. Let $s_i \in S_i$ be a pure strategy for player i, and $S_i(q) = \{s_i \mid s_i \in S_i, q \text{ is reached and taken in } s_i\}$. For $q' \in succ_i(q)$, $x_i(q',t) = \sum_{s_i\in S_i(q')} \rho(s_i,t)$. We need to show that this is also true at time $t+1$. By applying Equation 3 and standard replicator dynamics, this becomes

$$x_i(q',t)\frac{u_i(\mathbf{x}_{i\to g_{q'}})}{u(\mathbf{x})} = \sum_{s_i\in S_i(q')}\left(\rho(s_i,t)\frac{u_i(\rho_{i\to s_i})}{u(\rho)}\right)$$

We know that $u(\mathbf{x}) = u(\rho)$ by the statement of the theorem, so we can remove these denominators. Then, similarly to above the utilities can be decomposed into a dot product of vectors and re-arranged, leading to

$$\mathbf{u}(\mathbf{x}_{-i})\cdot(x_i(q',t)(\mathbf{g}_{q'}(\mathbf{x}_i)) = \mathbf{u}(\boldsymbol{\rho}_{-i})\cdot\sum_{s_i\in S_i(q')}\sigma(s_i,t)\boldsymbol{\rho}_{i\to s_i}.$$

The vectors on left are realization equivalent by the statement of the theorem. So, this is only true if the vectors on the right are realization equivalent. The right-side vectors are realization equivalent by [6, Lemma 9] since there is no dependence on the opponents nor number of players. □

Theorem 3. *Given a valid realization profile $\mathbf{x}(t) = (\mathbf{x}_1(t),\ldots,\mathbf{x}_n(t))$, a new $\mathbf{x}(t+\Delta t)$ produced by Equation 4 satisfies the sequence-form constraints.*

Theorem 4. *Given some game G with player set N, let $\rho(t)$ be a mixed strategy profile and $\rho(t+\Delta t)$ be the profile produced by standard continuous-time replicator dynamics using G's normal-form representation. Let $\mathbf{x}(t)$ be a realization profile and $\mathbf{x}(t+\Delta t)$ a realization profile in G's sequence-form produced by Equation 4. Then for all players $i \in N$, $\mathbf{x}(t+\Delta t)$ and $\rho(t+\Delta t)$ are realization equivalent.*

The proofs of Theorems 3 and 4 are identical to the proofs for the discrete case except $g_q(\mathbf{x}_i)$ is replaced by $\Delta g_q(\mathbf{x}_i)$.

3.2 Link to CFR Minimization

In this section, we show that discrete-time SFRD (equation 3) corresponds to a form of counterfactual regret minimization. As a result, under mild conditions the average strategies ($\bar{\mathbf{x}}$) converge to an equilibrium in two-player zero-sum games, as CFR minimization does by producing ϵ-equilibria with decreasing upper bounds on ϵ as the number of iterations increase. Similar to [12], we relate Polynomial Weights (PW) and replicator dynamics; however, unlike [12] we show that SFRD corresponds to regret minimization rather than analyzing the evolutionary dynamics of PW. In our case, SFRD corresponds to counterfactual regret minimization where PW replaces regret matching as the underlying no-regret learner at each information set.

To do this, we assign at every $I \in \mathcal{I}_i$ its own individual no-regret learner, denoted PW(I). Denote $w_{I,a}(t)$ as the weight of PW(I) at I for action $a \in A(I)$. The initial weights $w_{I,a}(1) = 1$ for all $(I,a) \in Q_i$. This leads to a fully-mixed behavioral strategy where at each I, player i chooses an action $a \in A(I)$ uniformly at random. Let $q = (I,a)$, we construct a loss for PW(I) with the following form,

$$l_{I,a} = \frac{\max_{a'\in A(I)} u_i(\mathbf{x}_{i\to g_{(I,a')}}) - u_i(\mathbf{x}_{i\to g_q})}{\Delta_{-i}(I)}, \text{ where} \quad (8)$$

$\Delta_{-i}(I) = \max_{a',a''\in A(I)}(u_i(\mathbf{x}_{i\to g_{(I,a')}}) - u_i(\mathbf{x}_{i\to g_{(I,a'')}}))$ is the counterfactual payoff range of I. It is easy to see that $l_{I,a} \in [0,1]$: in the numerator and denominator, the payoffs for all the terms such that $(h,z) \notin Z_I$ cancel out, leaving the payoffs when reaching I. The denominator is the largest value that the numerator can equal.

Lemma 1. *When using discrete-time SFRD, the update for $q = (I,a)$ from Equation 3 corresponds to an equivalent update of PW(I) with the loss $l_{I,a}$ as defined in Equation 8.*

PROOF. From Equation 3, let $w_i(t) = x_i(q,t)$, we have

$$\begin{aligned}
& w_i(t+1) = w_i(t)u_i(\mathbf{x}_{i\to g_q})/u_i(\mathbf{x}) \\
\Rightarrow\quad & w_i(t+1) = w_i(t)u_i(\mathbf{x}_{i\to g_q}) \\
\Rightarrow\quad & w_i(t+1) = w_i(t)(\max_{a'\in A(I)} u_i(\mathbf{x}_{i\to g_{(I,a')}}) - \Delta_{-i}(I)l_{I,a}) \\
\Rightarrow\quad & w_i(t+1) = w_i(t)(1 - \frac{\Delta_{-i}(I)}{\max_{a'\in A(I)} u_i(\mathbf{x}_{i\to g_{(I,a')}})} l_{I,a}).
\end{aligned}$$

The second and fourth lines follow because scaling by a constant at I does not affect the distribution at I due to normalization. The third line substitutes $u_i(\mathbf{x}_{i\to g_q})$ from Eq. 8. Here, $\eta = \Delta_{-i}(I)/(\max_{a'\in A(I)} u_i(\mathbf{x}_{i\to g_{(I,a')}}))$ is the learning rate of PW(I), and $0 \le \eta \le 1$ since $\Delta_{-i}(I)$ is a payoff range, subtracts the minimum value at I, and because $\max_{a'\in A(I)} u_i(\mathbf{x}_{i\to g_{(I,a')}})$ includes terms for $(h,z) \notin Z_I$. □

We now show that these individual losses are, in fact, forms of counterfactual regrets. First, we relate $g_q(\mathbf{x}_i)$ to counterfactual values.

LEMMA 2. *Given a realization profile $\mathbf{x}_i$ for player $i \in N$ and an equivalent behavioral strategy σ_i, let $q = (I,a)$, using $v_i(I,\sigma,a)$ as defined in Eq. 7 and for any $\sigma_{-i} \in \Sigma_{-i}$,*

$$\max_{a'\in A(I)} u_i(\mathbf{x}_{i\to g_{(I,a')}}) - u_i(\mathbf{x}_{i\to g_q}) = \max_{a'\in A(I)} v_i(I,\sigma,a') - v_i(I,\sigma,a).$$

PROOF. First, since $q = (I,a)$ and both $a, a' \in A(I)$, observe that the realization weights for profiles $\mathbf{x}_{i\to g(I,a')}$ and $\mathbf{x}_{i\to g_q}$ are identical for all $q' \in Q_i$ leading to histories outside Z_I. Therefore, the left side can be rewritten as

$$\begin{aligned}
& \max_{a'\in A(I)} \sum_{(ha',z)\in Z_I} \pi_i^\sigma(h)\pi_i^\sigma(ha,z)\pi_{-i}^\sigma(z)u_i(z) \\
& \quad - \sum_{(h,z)\in Z_I} \pi_i^\sigma(h)\pi_i^\sigma(ha,z)\pi_{-i}^\sigma(z)u_i(z) \\
= \; & \max_{a'\in A(I)} \sum_{(ha',z)\in Z_I} \pi_i^\sigma(ha,z)\pi_{-i}^\sigma(z)u_i(z) \\
& \quad - \sum_{(h,z)\in Z_I} \pi_i^\sigma(ha,z)\pi_{-i}^\sigma(z)u_i(z) \\
= \; & \max_{a'\in A(I)} v_i(I,\sigma,a') - v_i(I,\sigma,a).
\end{aligned}$$

The second line follows from the fact that $\pi_i^\sigma(h) = 1$ when i uses $g_q(\mathbf{x}_i)$ for $h \in I$, and distributions $\sigma(I')$ for I' that come after I are unchanged by Eq. 5. The last line follows from the definition of counterfactual value in Eq. 7. □

Define the **average immediate counterfactual regret** for player i at time T and information set I as in [26]:

$$\bar{R}_{i,imm}^T(I) = \frac{1}{T}\max_{a\in A(I)}\left(\sum_{t=1}^{T} v_i(I,\sigma^t,a) - v_i(I,\sigma^t)\right), \quad (9)$$

where σ^t is the profile used by both players at time t, and $v_i(I,\sigma^t) = \sum_{a\in A(I)} \sigma^t(I,a)v_i(I,\sigma^t,a)$ is the counterfactual value for playing $\sigma(I)$ at I.

As in the original work, we show that the combination of PW(I) learners over all I minimizes overall regret.

LEMMA 3. *Let R_i^T be overall regret for player i, and $\bar{R}_i^T = R^T/T$, and $(x)^+ = \max(0,x)$. The overall regret of discrete-time SFRD for player i is bounded by:*

$$\bar{R}_i^T \le \frac{\Delta_i}{T}\sum_{I\in\mathcal{I}_i}\left(\max_{a\in A(I)}\sum_{t=1}^{T}(l_{I,\sigma}^t - l_{I,a}^t)\right)^+,$$

where $\Delta_i \le \max_{z,z'\in Z}(u_i(z) - u_i(z'))$ is the payoff range, and $l_{I,\sigma}^t = \sum_{a\in A(I)} \sigma(I,a)l_{I,a}^t$.

PROOF. For $I \in \mathcal{I}_i$, $v_i(I,\sigma,a) - v_i(I,\sigma)$

$$\begin{aligned}
= \; & (\max_{a'\in A(I)} v(I,\sigma,a') - \Delta_{-i}(I)l_{I,a}) \\
& - \sum_{a''\in A(I)} \sigma(I,a'')\left(\max_{a'\in A(I)} v_i(I,\sigma,a') - \Delta_{-i}(I)l_{I,a''}\right) \\
= \; & -\Delta_{-i}(I)l_{I,a} + \sum_{a''\in A(I)} \sigma_i(I,a'')\Delta_{-i}(I)l_{I,a''} \\
= \; & \Delta_{-i}(I)(l_{I,\sigma} - l_{I,a}).
\end{aligned}$$

The first line follows from Eq 8, Lemma 2, and the definition of $v_i(I,\sigma)$ from above. The second line follows since $\sum_{a''\in A(I)} \sigma_i(I,a'') = 1$. Substituting into Eq. 9 leads to

$$\bar{R}_{i,imm}^T(I) = \frac{1}{T}\max_{a\in A(I)}\left(\sum_{t=1}^{T}\Delta_{-i}^t(I)(l_{I,\sigma}^t - l_{I,a}^t)\right). \quad (10)$$

Since $l_{I,\sigma}^t$, $l_{I,a}^t$ are bounded and due to perfect recall, the rest follows the proof of [26, Theorem 3], except with utilities replaced by bounded losses. □

There is one more small step before we present our main theorem. The standard Polynomial Weights algorithm uses a fixed learning rate η, whereas in SFRD the parameters of PW(I) may be different each iteration.

LEMMA 4. *Suppose a modified PW algorithm, PW', is used such with update rule $w(t+1) = w(t)(1-\eta^t l^t)$, and $\eta^t \le \frac{1}{2}$ for all $1 \le t \le T$. Then there exists some $\eta^* \le \frac{1}{2}$ such that regret of PW' at time T satisfies $R^T \le \eta^* Q_{min}^T + \ln K/\eta^*$.*

PROOF. Following the reasoning in [1, Theorem 4.6], the total weight $W^{T+1} = K\prod_{t=1}^{T}(1-\eta^t F^t)$, where F^t is the algorithm's loss at time t. For $0 \le \eta^t \le \frac{1}{2}$, there exists a fixed $\eta^* \in [0,\frac{1}{2}]$ s.t. $\prod_{t=1}^{T}(1-\eta^t F^t) = \prod_{t=1}^{T}(1-\eta^* F^t)$, so then $W^{T+1} = K\prod_{t=1}^{T}(1-\eta^* F^t)$ and the rest of the original PW bound analysis can be applied using the fixed η^*. □

We now present our main result.

THEOREM 5. *For player $i \in N$, if $\forall I \in \mathcal{I}_i$, $1 \le t \le T$,*

$$\eta_I = \frac{\Delta_{-i}^t(I)}{\max_{a'\in A(I)} u_i(\mathbf{x}_{i\to g_{(I,a')}}(t))} \le \frac{1}{2},$$

then overall regret of discrete-time SFRD is bounded by:

$$\bar{R}_i^T \le \frac{\Delta_i|\mathcal{I}_i|}{\eta^* T}\left(\ln|A_i| + (\eta^*)^2 Q_{min}\right),$$

where $Q_{min} = \max_{I\in\mathcal{I}_i} Q_{I,min}^T$, $|A_i| = \max_{I\in\mathcal{I}_i}|A(I)|$, and $\eta^* = \mathrm{argmax}_{\eta_I : I\in\mathcal{I}_i} Q_{I,min}^T$.

PROOF. Recall the general regret bound of PW from Section 2.2. By Lemma 4, each individual PW(I) is bounded by $R_{i,imm}^T(I) \le \eta_I^* Q_{I,min}^T + \ln|A(I)|/\eta_I^*$. Then, applying the bound to Equation 10, by Lemma 3 the overall bound is obtained by summing over all the information sets. □

The condition on the parameter η_I of Theorem 5 is satisfied for a large class of games and strategies. The numerator is a sum over a subtree. The denominator includes a sum over histories outside Z_I, therefore η will be very small in most cases. Even if the opponent concentrates all the weight to get to I, the sum includes terminal histories for which chance chose outcomes that do not lead to I.

The condition is not satisfied when there is a large gap between $\min_{a' \in A(I)} u_i(\mathbf{x}_{i \to g_{(I,a')}})$ and $\max_{a' \in A(I)} u_i(\mathbf{x}_{i \to g_{(I,a')}})$. This can happen, for example, in a perfect information game when the opponent, $-i$, concentrates their entire weight to reach a decision point $I \in \mathcal{I}_i$ where there is one action that leads to a certain loss for i and another to certain win for i.

4. EXPERIMENTS

In this section, we provide the first empirical evaluation of sequence-form replicator dynamics.

In our experiments, we use two separate metrics to evaluate the practical behavior of sequence-form replicator dynamics. The first is the observed rate of convergence to ϵ-Nash equilibrium. This is done by computing the smallest ϵ satisfying Equation 1. To compute ϵ for some profile σ in an n-player game, we first compute the expected values $u_i(\sigma)$ to each player. Then, for each player i we compute a best response to σ_{-i}, $\sigma_i^* \in \mathrm{BR}(\sigma_{-i})$, and $\epsilon_i = u_i(\sigma_i^*, \sigma_{-i}) - u_i(\sigma)$. Then, $\epsilon = \max_{i \in N} \epsilon_i$. The second metric is how well the resulting strategy σ performs against a fixed baseline player. Our baseline player chooses an action $a \in A(I)$ uniformly at random, i.e. with probability $1/|A(I)|$. Specifically, denote b_i as the baseline for player i, $b = (b_1, b_2, \ldots, b_n)$, and $u_i(b)$ as the expected value to player i if all players play the baseline strategies. The *overall performance* of some profile σ, is a measure of how much the players prefer to use their σ_i to play against the baseline opponents than b_i:

$$\text{Performance}(\sigma) = \sum_{i \in N} \left(u_i(\sigma_i, b_{-i}) - u_i(b) \right). \tag{11}$$

Under both evaluation metrics, we analyze the evolution of the *current strategy profile* at time t, $\mathbf{x}(t)$, as well as the *average strategy profile* up to time t, $\bar{\mathbf{x}}(t) = \frac{1}{t} \sum_{j=1}^{t} \mathbf{x}(j)$. When minimizing regret in two-player zero-sum games, the strategy that converges to an equilibrium is $\bar{\mathbf{x}}(t)$. Despite this, and unlike when $n = 2$, in multiplayer ($n > 2$) setting there is preliminary evidence that suggests using the current strategies works better, in practice, when minimizing standard counterfactual regret [7].

We focus on discrete-time replicator dynamics in a generalization of the well-studied game Kuhn poker.

4.1 Generalized Multiplayer Kuhn Poker

Kuhn poker is a simplified poker game originally proposed by Harold W. Kuhn [14]. Kuhn poker has been analyzed analytically by Kuhn and used as a testbed in studies in algorithms and multiagent systems [10, 21, 23].

Generalized n-player Kuhn poker, KP(n) consists of a deck with $n+1$ cards $\{1, 2, \cdots, n, n+1\}$. Every player starts with 2 chips. At the start of the game, each player antes one of their chips placing it into a central pot of $p = n$ chips. Then, chance deals one card to each player. The game starts with Player 1 and proceeds sequentially to the next player that is not eliminated. A player may bet (if they have a chip) or pass. To bet, a player puts the chip into the pot: $p \leftarrow p + 1$. A player can fold by passing when $n < p < 2r$ where r is the number of remaining players, *i.e.* players that have not been eliminated. If a player folds, they are eliminated and receive a payoff of -1 for the chip lost. The game proceeds until either (i) all but one player has folded, or (ii) $p = r$ and everyone has passed or all the remaining players have run out of chips. In the former case, the remaining player wins the pot. In the latter case, the remaining player with the highest-valued card wins the pot. In all cases, a player's payoff is the number of chips they end up with minus their 2 starting chips.

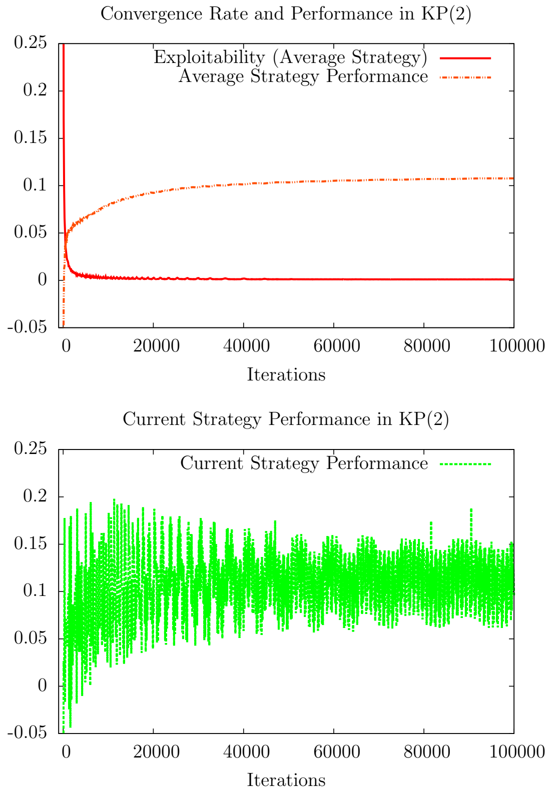

Figure 1: Top: convergence rate and overall performance of average strategy in KP(2). Bottom: overall performance of the current strategy. In both cases, the vertical axis represents utility.

The form of every equilibrium strategy for Kuhn's original game, KP(2) was derived analytically and used in a number of studies, *e.g.* [10]. The expected value to the first player in equilibrium is $-1/18$. In KP(3), an ϵ-equilibrium with $\epsilon = 0.0044563$ was found after 10^8 iterations of standard CFR minimization [21]. Recently, a family of parameterized equilibrium profiles was analytically derived for KP(3) [23]. We are unaware of any work studying KP(n) for $n > 3$.

In our implementation, a shifted payoff function $u_i'(z) = u_i(z) + 3$ is used to ensure that all the utilities are positive and that Equation 3 is well-defined. However, in our results we present utilities as described by the original game.

4.2 Two-Player Experiments

We first analyze the behavior of SFRD in two-player Kuhn poker. In two-player zero-sum games, a common way to compute distance to an equilibrium is to use exploitability. Since $u_2(\sigma) = -u_1(\sigma)$, the gap in Equation 1 is replaced by $\max_{\sigma_1' \in \Sigma_1} u_1(\sigma_1', \sigma_2) + \max_{\sigma_2' \in \Sigma_2} u_2(\sigma_1, \sigma_2')$. Therefore, in the case of two players, the graphs show exploitability.

The convergence rate and overall performance of the average and current strategy profiles are shown in Figure 1. The exploitability quickly drops to low values within the first few hundred iterations. The values for $t = (100, 200, 300)$ are $(0.916, 0.171, 0.096)$ to less than 0.001, with $u_1(\bar{\mathbf{x}}(t)) \approx -0.548$, $u_2(\bar{\mathbf{x}}(t)) \approx 0.558$ at $t = 100000$. The average performance slowly but steadily rises from -0.25 initially to 0.109 at $t = 100000$. The performance of the current strategy is much noisier, suggesting that players are switching significantly between alternatives, but slowly seems to concentrate to a neighborhood around 0.1 as well.

4.3 Multiplayer Experiments

In this section, we present experiments for KP(3) and KP(4). Results are shown in Figure 2.

In KP(3), we notice again that $\bar{\mathbf{x}}(t)$ reaches low values of ϵ relatively quickly, within a few thousand iterations, reaching $\epsilon \approx 0.00169$ at $t = 100000$. Interestingly, the current strategy profiles do not seem to reduce ϵ smoothly as the average strategy does. This could be because the evolution has not found any attracting stable fixed points or that the strategies are "orbiting" around a fixed point, possibly the one being approached by $\bar{\mathbf{x}}(t)$. We investigated further by comparing the strategies to the known equilibrium set described in [23]. Specifically, we computed a Euclidean distance between the particular probabilities $\sigma(I, a)$ necessary for all players to minimize their worst-case payoff, for points $t = 10^1, 10^2, 10^3, 10^4, 99980, 99990, 10^5$. These distances for the average profile are $(1.288, 0.642, 0.189, 0.026, < 0.003, < 0.003, < 0.003)$ and for the current profile are (1.152, 0.364, 0.312, 0.0002, $< 10^{-5}, < 10^{-5}, < 10^{-5}$). Each player's current strategy could be switching toward different approximate equilibrium strategies which when combined lead to worst-case penalties, while staying close to the important values that characterize the equilibrium. In contrast, the average strategies seem to focus on reducing ϵ.

From the performance graphs, we see that early on the performance spikes and and then slowly decreases. This could be due the players cooperating to increase the score against the early strategies close to the baseline, which slowly decreases as each player learns to counter the opponents, eventually stabilizing around 0.615. The current strategy performance graph is noisy as in KP(2), and the performance of the average strategy seems to be higher at first ($t \leq 30000$).

In KP(4), again the average profile seems to be reducing ϵ smoothly and getting closer to a Nash equilibrium over time, reaching $\epsilon \approx 0.0093$ at $t = 10000$. Like in KP(3), the ϵ convergence of the current strategies $\mathbf{x}(t)$ is erratic. In KP(4), the performance of the current strategy seems to be less than the average strategy $\bar{\mathbf{x}}(t)$, which is consistent with the KP(3) results.

In all cases, the average strategy profile $\bar{\mathbf{x}}$ appears to be converging to equilibrium with ϵ decreasing as the number of iterations increase. Since the dynamics are realization-equivalent to standard replicator dynamics, dominated strategies will not survive in the current strategies and will slowly be played less and less in the average strategies as well.[1] In all cases, both current and average strategy profiles perform better than the baseline, but the average strategies sometimes perform better and fluctuate less over time.

5. CONCLUSION AND FUTURE WORK

In this paper, we further develop sequence-form replicator dynamics. First, we have shown that SFRD do not violate sequence-form constraints and that SFRD updates are realization-equivalent to standard replicator dynamics for $n > 2$ players. Second, SFRD can minimize regret, leading to convergence guarantees in two-player zero-sum games. Finally, we have provided the first empirical evaluation of SFRD, showing convergence of average strategies to ϵ-Nash equilibria in Kuhn poker with two, three, and four players.

In future work, we aim to apply SFRD to larger, more complex games and further characterize the stability of the resulting strategies. In addition, we hope to provide conditions that lead to producing equilibrium refinements, such as sequential or trembling-hand perfect equilibria.

Acknowledgments. Thanks to Richard Gibson for answering questions about CFR, Kuhn poker, and dominated actions in the $(n > 2)$-player setting. This work is partially funded by the Netherlands Organisation for Scientific Research (NWO) in the framework of the project Go4Nature, grant number 612.000.938. This paper is dedicated to Richard Edgar Lanctot, who always believed in me.

6. REFERENCES

[1] A. Blum and Y. Mansour. Learning, regret minimization, and equilibria. In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*, chapter 4. Cambridge University Press, 2007.

[2] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

[3] R. Cressman. *Evolutionary Dynamics and Extensive Form Games*. MIT Press, Cambridge, MA, USA, 2003.

[4] J. Flesch, T. Parthasarathy, F. Thuijsman, and P. Uyttendaele. Evolutionary stochastic games. *Dynamic Games and Applications*, 3(2):207–219, 2013.

[5] D. Fudenberg and D. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

[6] N. Gatti, F. Panozzo, and M. Restelli. Efficient evolutionary dynamics with extensive-form games. In *27th AAAI Conference on Artificial Intelligence*, 2013.

[7] R. Gibson. Regret minimization in non-zero-sum games with applications to building champion multiplayer computer poker agents. *CoRR*, abs/1305.0034, 2013.

[8] H. Gintis. *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction*. Princeton University Press, second edition, 2009.

[9] D. Hennes, K. Tuyls, and M. Rauterberg. State-coupled replicator dynamics. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'09)*, 2009.

[10] B. Hoehn, F. Southey, R. Holte, and V. Bulitko. Effective short-term opponent exploitation in simplified poker. In *AAAI'05*, pages 783–788, 2005.

[1] SFRD may also filter out strictly dominated actions in the limit as $T \to \infty$ (as standard CFR minimization does, see [7]), but this remains an open research question.

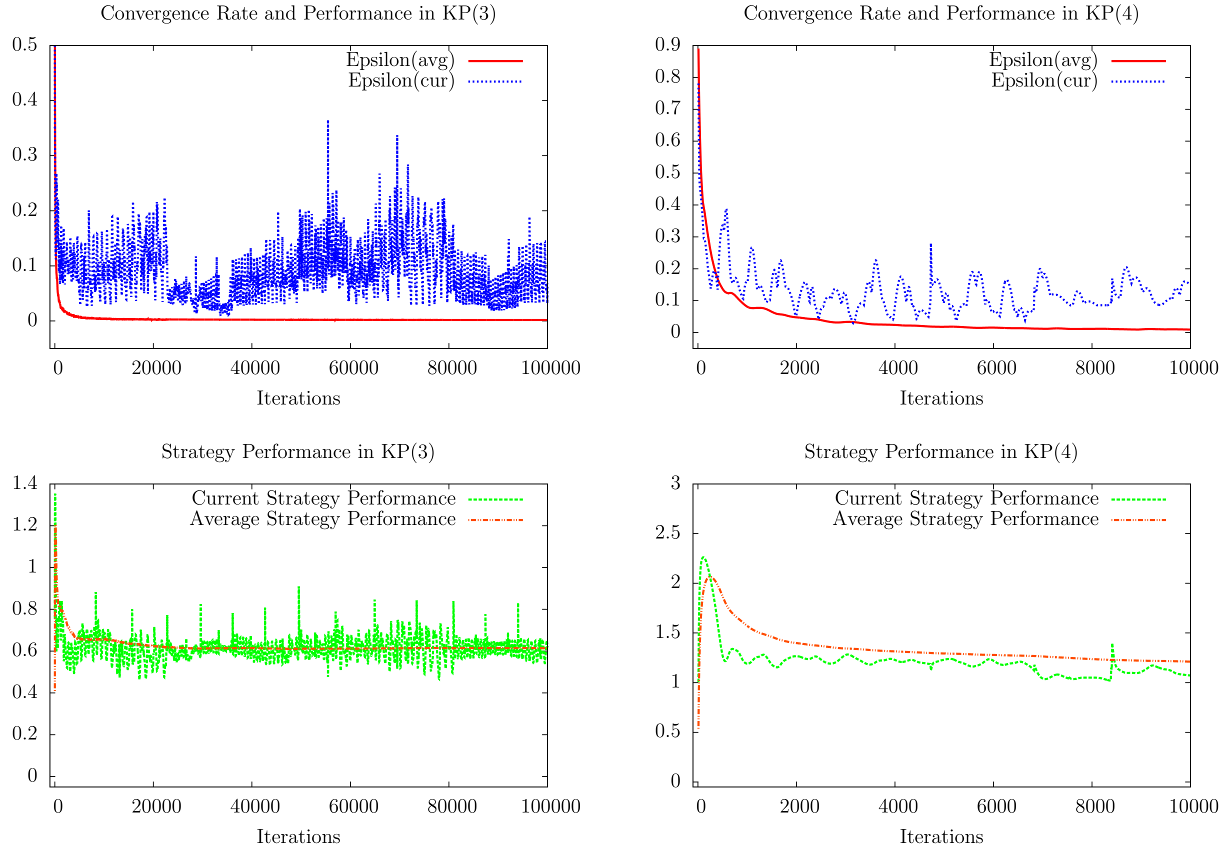

Figure 2: Convergence rate and overall performance in KP(3) (left) and KP(4) (right). The vertical axes represent utility. Epsilon(avg) and Epsilon(cur) represent the convergence rates of $\bar{\mathbf{x}}(t)$ and $\mathbf{x}(t)$, respectively.

[11] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[12] T. Klos, G. van Ahee, and K. Tuyls. Evolutionary dynamics of regret minimization. *Machine Learning and Knowledge Discovery in Databases*, 6322:82–96, 2010.

[13] D. Koller, N. Megiddo, and B. von Stengel. Fast algorithms for finding randomized strategies in game trees. In *26th ACM Symposium on Theory of Computing (STOC '94)*, pages 750–759, 1994.

[14] H. Kuhn. Simplified two-person poker. *Contributions to the Theory of Games*, 1:97–103, 1950.

[15] H. Kuhn. Extensive games and the problem of information. *Contributions to the Theory of Games*, 2:193–216, 1953.

[16] N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[17] J. Maynard-Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

[18] J. Maynard-Smith and G. Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.

[19] M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[20] M. Ponsen, K. Tuyls, M. Kaisers, and J. Ramon. An evolutionary game-theoretic analysis of poker strategies. *Entertainment Computing*, 1:39–45, 2009.

[21] N. A. Risk and D. Szafron. Using counterfactual regret minimization to create competitive multiplayer poker agents. In *AAMAS*, pages 159–166, 2010.

[22] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[23] D. Szafron, R. Gibson, and N. Sturtevant. A parameterized family of equilibrium profiles for three-player kuhn poker. In *Proceedings of the Twelfth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 247–254, 2013.

[24] P. Taylor and L. Jonker. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40:145–156, 1978.

[25] K. Tuyls and S. Parsons. What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence*, 171:406–416, 2007.

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