HETEROGENEITY, REINFORCEMENT LEARNING AND CHAOS IN POPULATION GAMES

JAKUB BIELAWSKI, THIPARAT CHOTIBUT, FRYDERYK FALNIOWSKI, MICHAŁ MISIUREWICZ, AND GEORGIOS PILIOURAS

Abstract. Traditional evolutionary game theory is a powerful tool for analyzing the statistics of a large population participating in a game. However, the behavior of the individual agents are based on simple memoryless dynamics and this collective behavior is typically represented by a single distribution encoding the frequency of the different actions played deterministically by all the infinitesimal agents. In this paper, we study a more general model that captures a large population of agents of different types, each of them performing reinforcement learning, leveraging memory of past actions' performance and outputting unpredictable behavior. The state of the system is captured not by a single discrete distribution but involves more complex measures capturing all possible heterogeneous learning states of the population of agents. We apply this advanced learning model in congestion games, which are well known to admit an essentially unique equilibrium solution. We showcase that our learning dynamics can exhibit convergence to numerous asymmetric equilibrium states as well as phase transitions to chaos. Remarkably, even in the chaotic regime, precise predictions can be made about the system performance as the time-average cost of all actions are shown to be equal to each other and in fact agree with their values at equilibrium. Therefore, a plethora of novel heterogeneous normative solutions are shown to be dynamically emergent in population games.

1. Introduction

Learning in games with a large population of agents has been a staple of evolutionary game theory [32, 16, 38]. In such models a large population of agents are presumed to choose one action/type from a fixed number of options. Each individual agent is assumed to update their action according to simple, memoryless models (e.g., each agent samples another one in the population uniformly at random and if the action played by the sampled agent outperforms their current action then the new action is chosen with some probability depending on the payoff improvement). At the population level, the behavioral model then examines the deterministic limit of this process given an infinitely large population and tracks how the frequencies of the different actions in the population vary with time using a single probability distribution.

Multi-agent reinforcement learning on the other hand is traditionally studied in games with only a handful of agents, see [5, 15, 39]. Each agent applies sophisticated learning by keeping explicit memory of how different actions have fared during past plays. Using this historical information each agent updates their beliefs about which action is best for them to play next and sample their future actions according to these beliefs. Unfortunately, even in games with only two agents reinforcement learning dynamics, such as the replicator dynamics [36, 34], can lead to chaotic behavior

(in the sense of positive Lyapunov exponents of the system) even in rather simple two-agent games, such as slight generalizations of the standard Rock-Paper-Scissors game [33]. These results have direct analogues in evolutionary game theory where chaos can emerge even with only four strategies/types [35]. Furthermore, such chaotic results are common in two-players games [13, 26, 8, 24, 10, 1] as well as multi-player games [31].

Moving towards combining reinforcement learning and population games, [9, 4] proved phase transitions from global stability to Li-Yorke chaos in population games, known as non-atomic congestion games [25]. The learning dynamics corresponds to discrete-time variants of replicator dynamics, which allows for the population as a whole to learn/adapt either at a slow or fast pace. When the learning rate is small, the behavior is qualitatively similar to continuous-time models which are known to converge to an essentially unique Nash equilbrium/flow due to a potential/Lyapunov function argument [18, 19]. In contrast, when the learning rate increases, chaos emerges, unlike in the continuous-time variant. Critically, however, these models only allow for homogeneous behaviors amongst a continuum of agents. In the orthogonal direction, [20] studied reinforcement learning in a population game, where the population statethe probability distribution over the set of mixed strategiesevolves according to a partial differential equation variant of replicator dynamics. Although this model allows for heterogeneity of strategies to persist even at equilibrium, no instability results are reported. Instead, in specific classes of potential games, the population mean (i.e. the expected mixed strategy over the whole population) converges to the set of Nash equilibria.

In this work, we combine essential features of these two models by explicitly allowing for a (possibly infinitely) large population of agents each of which adapts its behavior according to standard class of online/reinforcement learning dynamics known as Multiplicative Weights Updates (MWU).[1] Moreover, our model to the best of our knowledge is the first to combine two distinct sources of heterogeneity amongst the agents. First, each agent may start with different beliefs about which is the most promising action to play initially. Thus, the (evolving) state of the system will be intrinsically complex as it encodes a function from (possibly a continuum of) agents/types to a continuum of probability distributions/beliefs. Second, in our discrete-time dynamics each agent has their own intensity of learning (learning rates), i.e., some agents may adapt quickly to payoff signals by learning fast whereas others may be much more patient in how they update their beliefs. Given this complex learning model we are interested in understanding under what conditions and to

[1]MWU is a ubiquitous online learning meta-algorithm that, along with several close variants thereof, has been rediscovered several times under different names. First introduced by Hannan [14], then rediscovered from 1990's with Aggregating Algorithm (AA) by Vovk [37], Smooth Fictitious Play by Fudenberg and Levine [12], Weighted Majority Algorithm by Littlestone and Warmuth [22], Hedge by Auer et al. [3], Multiplicative Weights by Freund and Schapire [11], Exponentially Weighted Average Forecaster by Cesa-Bianchi and Lugosi [6] the Exponentiated Gradient algorithm (EG) by Kivinen and Warmuth [17] and the Discrete Replicator Equation by Losert and Akin [23], which itself is closely connected to many other models of natural/evolutionary selection (see [23, 7] for more discussion). It is arguably one of the most well studied dynamics in game theory [6, 29]. For the more detailed history of MWU, see [7, 2].

what extent does the resulting learning behavior agree with game theoretic solution concepts such as Nash equilibria.

We employ this behavioral model in a standard game theoretic setting of congestion (potential) games and show phase transitions where the system can diverge from global stability to instability and chaos. Interestingly, even when the system is convergent the system has a continuum of equilibria. Nevertheless, in this case we show that learning converges pointwise to a single equilibrium with the initial condition-dependent equilibrium selection process. Finally, when cost functions of different strategies differ, aggressive behavior of agents will inevitably lead to chaotic, complex behavior of the system with periodic orbits of every period as well as sensitivity to initial conditions (buttefly effects). Despite the chaotic evolution of the day-to-day behavior of the system, the *time-average costs* of the strategies of the agents as well as *time-average total flow* converge to their equilibrium values defined in the standard game theoretic setting. Thus, macroscopic order and regularity predicted by static game theory can emerge from persistent chaotic learning dynamics in the microscopic level.

2. Model

First, let us introduce the framework of game theory. A game has several/infinitely many players (agents), and each player has a set of possible strategies (actions). Each agent has a utility/cost, capturing the way in which his strategies and the strategies of other agents affect this agent's well-being. In this note, we consider a nonatomic congestion game with continuum agents and with only two possible (pure) strategies. Congestion games is a class of games introduced by Rosenthal [28]. In a congestion game, agents are choosing strategies (resources/paths), and the cost of the strategy depends on the total amount of agents choosing the same strategy. One of the most well studied applications of this class of games is in the modeling of traffic jams. In a nonatomic game, each agent controls an infinitesimally small fraction of the flow [25, 19]. Thus, there doesn't exist an (individual) agent whose change of behavior will affect the outcome of the game. Finally, a game-theoretic solution of a game is Nash equilibrium, a strategy profile in which all agents use their best response actions, thus none of them has an incentive to change their behavior. In the case of non-atomic congestion games such outcomes are also referred to as equilibrium flows or Wardrop equilibria.

Agents in our game update their beliefs according to Multiplicative Weights Update rule. How do the agents choose and adjust their choice of strategies over repeated play? Each (type of) agent may be seen as selecting its new distribution over strategies at each day/round so as to minimize a certain convex combination of cumulative costs and the Shannon entropy of its distribution over actions.

We will start by considering a simplified model with a finite number m of different types. With each type one can assign beliefs and learning rates. Out of the total flow/demand N, the group of agents of type i has size Nc_i.

We will denote the fraction of the number of the players of type i using the first strategy by x_i. The second strategy is chosen by $1 - x_i$ fraction of the number of the players of type i.

We will assume that the cost is proportional to the *load*. If we denote by $C(j)$ the total cost of players playing the strategy number j, and the coefficients of proportionality are α, β, then we get

$$C(1) = \alpha N \sum_{i=1}^{m} c_i x_i, \quad C(2) = \beta N \sum_{i=1}^{m} c_i (1 - x_i). \tag{1}$$

For the Multiplicative Weights Update, there are parameters $\varepsilon_i \in (0, 1)$, which can be treated as the common learning rates of all players of type i. Then the updated ratio will be (see SI text)

$$\frac{x_i(1-\varepsilon_i)^{C(1)}}{x_i(1-\varepsilon_i)^{C(1)} + (1-x_i)(1-\varepsilon_i)^{C(2)}}. \tag{2}$$

By combining formulas (1) and (2), we get

$$\frac{x_i}{x_i + (1-x_i)(1-\varepsilon_i)^{N(\beta \sum c_j - (\alpha+\beta)\sum c_j x_j)}}. \tag{3}$$

Then, taking $a_i = N(\alpha+\beta)\log\frac{1}{1-\varepsilon_i}$ and $b = \frac{\beta}{\alpha+\beta}$, and observing that $\sum c_j = 1$, we can rewrite formula (3) as[2]

$$\frac{x_i}{x_i + (1-x_i)\exp(a_i(\sum_{j=1}^{m} c_j x_j - b))}. \tag{4}$$

As ε_i increases if and only if a_i increases, we will refer to a_i as the learning rate[3], while b will describe the asymmetry of costs of the two strategies. By (4) the dynamics of types is described by the map

$$\begin{aligned} F(x_1, \ldots, x_m) = \Big(& \frac{x_1}{x_1 + (1-x_1)\exp(a_1(\sum_{j=1}^{m} c_j x_j - b))}, \\ & \ldots, \frac{x_m}{x_m + (1-x_m)\exp(a_m(\sum_{j=1}^{m} c_j x_j - b))} \Big). \end{aligned} \tag{5}$$

In this paper we are going to study long-term dynamics for any number of types and its dependence on how aggressively agents behave. In fact, we even do not have to assume that the number of types is finite. Namely, we can reinterpret our model in the following way.

We consider a space Ω with the probability measure μ, describing types of agents. Assignment of frequency of using the first strategy by a given type of agents (to which belief is assigned) will be done by a measurable function $\zeta\colon \Omega \mapsto (0, 1)$, while the learning rate for each agent will be given by $a\colon \Omega \mapsto (0, \infty)$. Thus, in the

[2]to simplify exposition and algebra, but there is no reason that it is essential for the results. In fact Scaling N is mathematically equivalent to scaling a_i's. Although we will not pursue this case, one can look at results for aggressive agents as the result on the consequences of increasing the demand of the system. For a more thorough discussion of consequences of taking this perspective see [9].

[3]There is another interpretation of a_i which can be found in the literature — intensity of choice. The larger a_i is the more important for agent to optimize his behavior based on the information received until now.

case considered above we have $\Omega = \{1, 2 \ldots, m\}$, $\mu(\{i\}) = c_i$, $a(i) = a_i\omega \in \Omega$ the formula (5) becomes

$$F(\zeta)(\omega) = \frac{\zeta(\omega)}{\zeta(\omega) + (1 - \zeta(\omega)) \exp\left(a(\omega)\left(\int \zeta \, d\mu - b\right)\right)}. \tag{6}$$

Thus, the general model will consist of a space Ω with a probability measure μ, measurable functions $\zeta\colon \Omega \mapsto (0,1)$, $a\colon \Omega \mapsto (0,\infty)$, and a constant $b \in (0,1)$. Let $I = (0,1)$. Let $M(\Omega, I)$ be the space of measurable functions from Ω to I. Then we consider the operator $F\colon M(\Omega, I) \mapsto M(\Omega, I)$ defined by (6) and study its dynamics.[4]

Before the analysis is performed, we first provide the necessary ingredients to rigorously understand the chaotic behaviors in our system.

Definition 1 (Li-Yorke chaos)**.** *Let (X, f) be a dynamical system and $x, y \in X$. We say that (x, y) is a* Li-Yorke pair *if*

$$\liminf_{n\to\infty} dist(f^n(x), f^n(y)) = 0,$$

and

$$\limsup_{n\to\infty} dist(f^n(x), f^n(y)) > 0.$$

A dynamical system (X, f) is Li-Yorke chaotic *if there is an uncountable set $S \subset X$ (called* scrambled set*) such that every pair (x, y) with $x, y \in S$ and $x \neq y$ is a Li-Yorke pair.*

The origin of the definition of Li-Yorke chaos is in the seminal Li and Yorke's article [21]. Intuitively, the orbits of two points from the scrambled set approach each other arbitrarily close and then go far from each other infinitely many times but (if X is compact) it cannot happen simultaneously for each pair of points. Why should a system with this property be chaotic? Obviously the existence of a large scrambled set implies that orbits of points behave in unpredictable, complex way. More arguments come from the theory of interval transformations, in the context in which it was introduced. For such maps the existence of one Li-Yorke pair implies the existence of an uncountable scrambled set and it is not very far from implying all other properties that have been called chaotic in this context, see e.g. [30]. In general, Li-Yorke chaos has been proved to be a necessary condition for many other "chaotic" properties to hold.

3. Analysis

The map $F(\zeta)$ describes how choices of agents change. As fixing ζ assigns to the type of agents a ratio of agents of this type choosing the first strategy (path/resource), $F(\zeta)(\omega)$ describes probability of choosing the first strategy in the next stage. Operator F is continuous on $M(\Omega, I)$ in the topology of pointwise convergence, since I is bounded, μ is finite, and the integral of ζ is a continuous function of ζ.

Topological conjugacy. The space of measurable functions is a complicated multidimensional object. But instead of studying the dynamics of F from (6) one can analyze the dynamics introduced by a map of the real line. To this aim we use

[4]While the formalism with the space (Ω, μ) and functions ζ is simple to work from the mathematical point of view, one can also think in different terms. Namely, instead of the function ζ, one can consider the measure $\zeta_*(\mu)$ on I (given by $\zeta_*(\mu)(X) = \mu(\zeta^{-1}(X))$.

topological conjugacy. Fix $\xi \in M(\Omega, I)$[5] and define a one-parameter family $(\xi_s)_{s\in\mathbb{R}}$ of elements of $M(\Omega, I)$ by

$$\xi_s(\omega) = \frac{\xi(\omega)}{\xi(\omega) + (1-\xi(\omega))\exp(sa(\omega))}.$$

It is easy to check that $\xi_0 = \xi$. Moreover, the function $s \mapsto \xi_s$ is continuous and strictly decreasing. Therefore, it is a homeomorphism from $\mathbb{R}$ onto $(\xi_s)_{s\in\mathbb{R}}$.

Lemma 3.1. *Any fixed $\xi \in M(\Omega, I)$ can be embedded in a one-parameter family, invariant for F, on which F is topologically conjugate to the map of the real line. That is, we have*

$$F(\xi_s)(\omega) = \xi_{G(s)}(\omega)$$

where

$$G(s) = s + \int \xi_s \, d\mu - b. \tag{7}$$

Lemma 3.1 implies once the assignment of using first strategy by each type of agent is chosen (ξ is fixed), the game dynamics can be studied by looking at the dynamics introduced by (7). Then, instead of working with the operator defined on the multidimensional space (usually of infinite dimension), one can study dynamics of the one-dimensional map. Due to the conjugacy, results for the map of the real line can be applied to the set on which $F(\xi)$ evolves.[6]

Dependence on learning rate. In this note we analyze what happens for varied values of the learning rate. To stress the dependence of G and ξ_s on the learning rate, we will write G_a instead of G, F_a instead of F and $\xi_{a,s}$ instead of ξ_s. The operator F_a has multiple fixed points, usually infinitely many. Nevertheless, once ξ is fixed (and thus G_a is introduced by (7)), the equilibrium is unique.

Theorem 3.2. *For every learning rate a the map G_a has a unique fixed point s_a^*. Moreover, s_a^* is a fixed point of G_a if and only if the costs of both strategies are equal.*

From Theorem 3.2 and Lemma 3.1 follows uniqueness of the fixed point once $\xi \in M(\Omega, I)$ is chosen. As the costs of both strategies are equal, no individual is motivated to change strategy, thus s_a^* determines Nash equilibrium of the game. So, knowledge of initial assignment of choosing strategies by every type of agents dictates unique equilibrium state, which is Nash equilibrium of the game. Therefore, usually the game has infinitely many Nash equilibria.

Remark 3.3. *From (7) we see that $G_a(s) = s$ only when $\int \xi_s d\mu = b$, so the expected value of ξ_s is equal to the asymmetry of costs ratio b.*[7]

Corollary 3.4. *If ζ is a fixed point of F_a then ζ is a Nash equilibrium of the game.*

[5]Throughout the paper, when we refer to an arbitrary function from $M(\Omega, I)$ we denote it by ζ, while once we use a fixed function we denote it by ξ.

[6]The set on which $F(\xi)$ evolves is usually complicated, see e.g. Figure 1.

[7]One may want to compare heterogeneous with homogeneous case. In the latter, see [9], the only Nash equilibrium is b, which attracts everything for sufficiently small a. As we show in the examples this is not the case for heterogeneous case.

All these equilibria share two common features: costs of both strategies at equilibrium are equal and the expected value of ξ_s is equal to the asymmetry of costs ratio.

We introduce two examples, which we will use further in the article to concretely demonstrate our results. The first one describes a setting in which a change in the learning rate of agents is induced by a change of one parameter. This is possibly the simplest setting that encapsulates all the complex phenomena studied in this work.

Example 1. *We introduce a one parameter family of maps. Namely, we fix the function a and consider the family of maps $\{G_{Aa}\}_{A\in(0,\infty)}$. We will call it an A-family. Let us look at a concrete simple (but non-trivial) A-family. We consider two types of agents $\Omega = \{1, 2\}$ with measure μ equally distributed $\mu(\{1\}) = \mu(\{2\}) = 0.5$. The learning rates are $a(1) = 1$, $a(2) = 3$, the asymmetry of costs is set $b = 0.3$ and we choose the map ξ such that $\xi(1) = 0.2$ and $\xi(2) = 0.6$. Then*

$$\xi_{Aa,s}(1) = \frac{0.2}{0.2 + 0.8\exp(As)} = \frac{1}{1 + 4\exp(As)},$$

and

$$\xi_{Aa,s}(2) = \frac{0.6}{0.6 + 0.4\exp(3As)} = \frac{3}{3 + 2\exp(3As)}.$$

Therefore,

$$G_{Aa}(s) = s + \frac{1}{2 + 8\exp(As)} + \frac{3}{6 + 4\exp(3As)} - 0.3.$$

In the second example, we consider a more heterogeneous setting consisting of multiple (999) types of agents with equally distributed measure. This example demonstrates a more realistic learning dynamics of a diverse population.

Example 2. *Let $\Omega = \{1, 2, 3, \dots, 999\}$ with equidistributed measure. The learning rate parameters for each type of agents are defined by $a(i) = 1.2(5 + (i \bmod 31))$. The assymetry of costs is set to $b = 0.3$. The initial condition is such that $\zeta(i) = 0.0001 + (i \bmod 23)/24$ with $s = 0$.*[8]

Learning in games can provide the basis for equilibrium prediction, which is exceptionally desirable e.g. in economics or computer science, as convergence to equilibrium guarantees predictable long-term behavior. From this perspective, fundamental question is whether the behavior of the system will stabilize at the static equilibrium prediction (Nash equilibrium). We show that this depends on the (choice of) a. In fact, learning dynamics converges exactly to one of the equilibria when agents are learning slowly. dynamics is relatively simple.

Theorem 3.5. *If $\sup_{\omega\in\Omega} a(\omega) < 8$, then fixed point of G_a is globally attracting.*

Corollary 3.6. *If $\sup_{\omega\in\Omega} a(\omega) < 8$, then for every $\zeta \in M(\Omega, I)$ the sequence $(F^n(\zeta))_{n=0}^{\infty}$ converges pointwise to a fixed point of F.*

Therefore, as long as learning rates of all agents are small, the system will equilibrate at the unique fixed point of G_a, where the costs of both paths are equal. Thus, the dynamics will be relatively simple and the equilibrium prediction agrees with the

[8] $\zeta_*(\mu)$ is a good approximation of the Lebesgue measure on the interval $[0, 1]$.

long-term behavior of learning in games. Nevertheless, the equilibrium will change due to the way in which the assignment of frequencies for types of agents is performed (the choice of ζ).

Now, let us consider the case when agents behave aggressively (more greedy algorithmically).

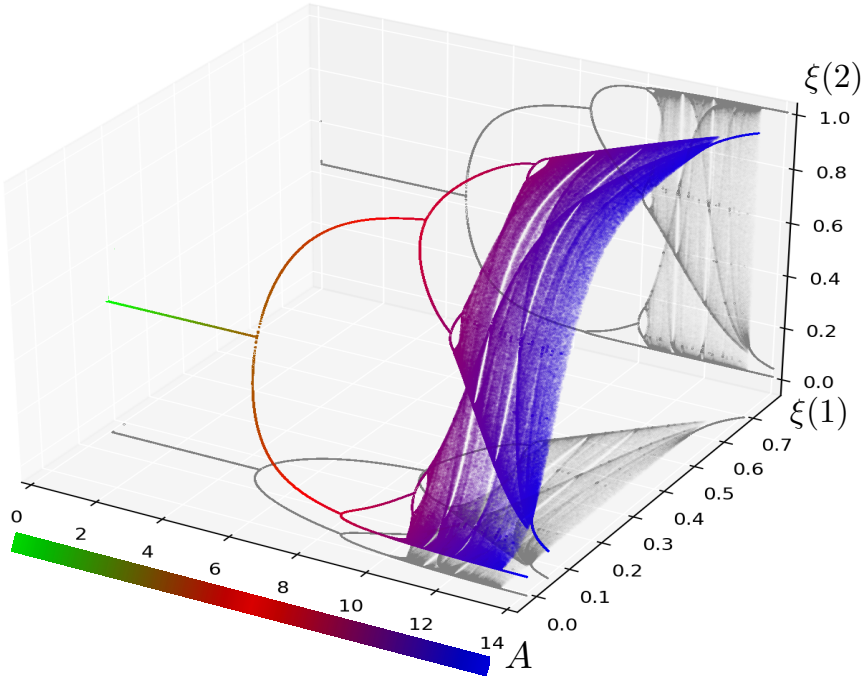

FIGURE 1. Bifurcation diagram of the whole system for Example 1. The diagram lives in a two-dimensional manifold embedded in a three-dimensional space. The shadows (in gray) are the projection of the flow onto $\xi(1)$ and $\xi(2)$ planes. Parameter A varies from 0 to 14. For low values of A the flow converges to a fixed point of the map G_{Aa}. As A increases, the flow becomes unstable or chaotic, in agreement with Theorem 3.7.

Theorem 3.7. *If $b \neq 1/2$ and the sequence $(a_k)_{k=1}^{\infty}$ of measurable functions from Ω to $(0,\infty)$ converges pointwise to infinity, then there exists K such that for every $k \geq K$ the map G_{a_k} has periodic orbits of all periods and is Li-Yorke chaotic.*

Corollary 3.8. *Let $b \neq 1/2$. For any choice of $\zeta \in M(\Omega, I)$ if the sequence $(a_k)_{k=1}^{\infty}$ of measurable functions from Ω to $(0,\infty)$ converges pointwise to infinity, then there exists K such that for every $k \geq K$ the map $F_{a_k}(\zeta)$ is Li-Yorke chaotic.*

Intuitively, the map is Li-Yorke chaotic if there exists a set of points such that orbits of any two distinct points from this set approach each other arbitrarily close and then go far from each other infinitely many times. Obviously the existence of such set implies that orbits of points behave in unpredictable, complex way. Theorem 3.7 and Corollary 3.8 imply that, when cost functions differ, if all types of agents behave aggressively enough (agents choose their strategies with sufficiently large learning

rate), then the system will inevitably become chaotic. In such case any long-term behavior will become extremely complex. We land in an unpredictable regime with periodic orbits of different periods, sensitive dependence on initial conditions and complicated dynamics. Yet, interestingly, time-average macroscopic order can emerge, in agreement with static game-theoretic predictions.

Time-average behavior and the equilibrium predictions. We see already that aggressive behavior of agents destabilizes the system. Nevertheless, once we look at the macroscopic level (total flow), from the perspective of time averages, the time-average total flow will eventually equilibrate at the game-theoretic equilibrium flow value. For a given $\zeta \in M(\Omega, I)$, we consider its space average, $\int \zeta \, d\mu$. This expected value of ζ can be interpreted as the total flow in strategy 1.

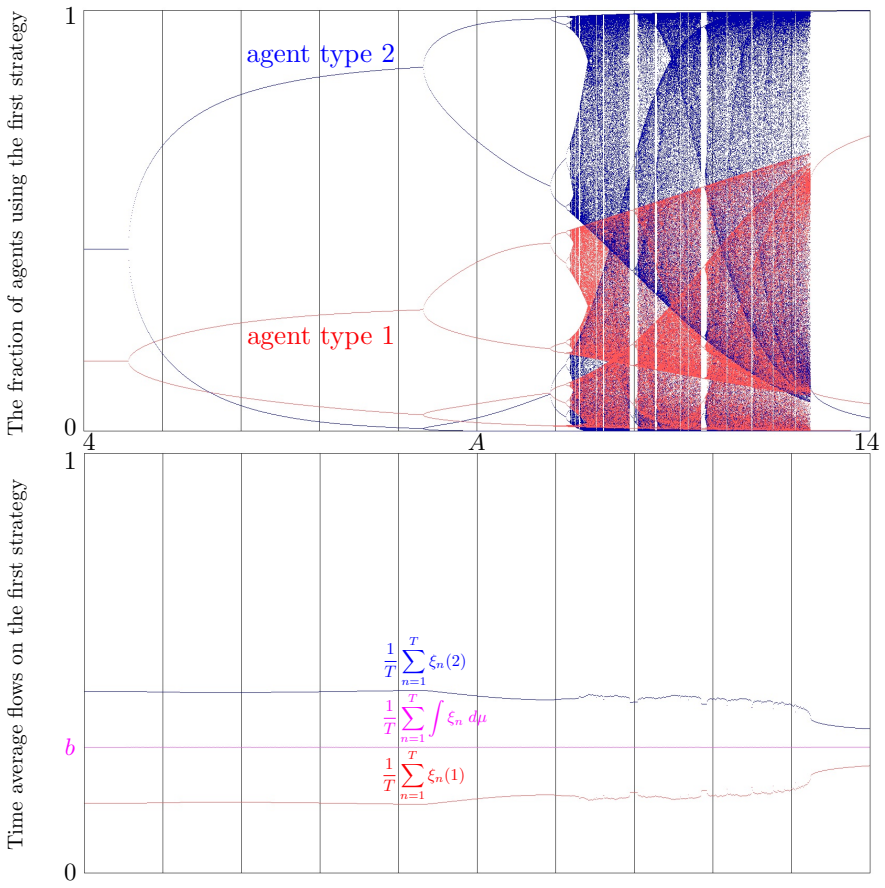

FIGURE 2. Illustration of Corollary 3.10 in the A-family of Example 1. (Top) Bifurcation diagrams for each type of agent as one varies A (the projection onto $\xi(1)$ and $\xi(2)$ of Figure 1). As A increases, the flow becomes unstable or chaotic. (Bottom) Despite instability or chaos, the time average *total* flow converges to the equilibrium flow b (magenta), but the time average flow of each agent type converges to a different value. Both plots are the results of the dynamics with $T = 10^3$ iterates, which are obtained after a burn-in period with 10^4 iterates.

Theorem 3.9. *If $\zeta_n = F^n(\zeta)$ then there exists $B \geqslant 0$ such that for every $T \geqslant 1$ we have*

$$\left| \sum_{n=0}^{T-1} \int \zeta_n \, d\mu - Tb \right| \leqslant B.$$

Corollary 3.10. *If $\zeta_n = F^n(\zeta)$ then*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{n=0}^{T-1} \int \zeta_n \, d\mu = b,$$

that is, the sequence of averages of the expected values of $F^n(\zeta)$ (which is the same as the sequence of expected values of averages of $F^n(\zeta)$) converges to b.

Corollary 3.10 tells us that the time averages of the space averages of the images of ζ converge. We can ask whether the same is true if we do not take space averages. What we know about the family of quadratic maps (and the bifurcation diagram) suggests that for almost every value of A for G_{Aa} either there is a globally attracting periodic orbit or an invariant measure absolutely continuous with respect to the Lebesgue measure. Thus, we can expect that, by the Birkhoff Ergodic Theorem, the time averages converge for almost every starting point. However, we cannot expect that the limit behaves nicely as a function of A (see Figures 2 and 3). Moreover, these figures show that for most of the values of A this limit is different than for the fixed point (by Theorem 3.5, the limit for the fixed point – which is the same as the value at the fixed point – is independent of A). In fact, if there exists an absolutely invariant measure and the limit almost everywhere for this measure is different than the value at the fixed point, then one can show that there are points for which the limit does not exist.

Thus, although behavior of individual trajectories can be complicated, their time-averages will always converge. Moreover, this convergence is to the same value independently on the choice of ζ and a. In addition, the time-average costs will also converge.

Corollary 3.11. *The average cost of each strategy converges to $Nb(1-b)$.*

Corollary 3.11 shows that time averages of both costs converge to the cost at the Nash equilibrium of a game b. Thus, the limiting cost is the same as the cost of the static game-theoretic prediction (cost at Nash equilibrium of the homogeneous game).

4. Discussion

Our reinforcement learning dynamics in games with heterogeneous populations defined by (6) contains infinitely many Nash equilibria ζ^* , each of them satisfying the total flow condition $\int \zeta^* d\mu = b$. However, these equilibria are attracting only when all the agents learn slowly (Corollary 3.6), then the learning dynamics will stabilize at one of the equilibria. With fast learning agents, the learning dynamics become unstable or chaotic (Corollary 3.8). Despite microscopic unpredictability of the learning dynamics, the macroscopic *time-average total flow* and the *time-average cost* converge to those defined by the total flow of value b (Corollary 3.10 and Corollary 3.11, respectively). In fact, b is the equilibrium flow in the homogeneous population

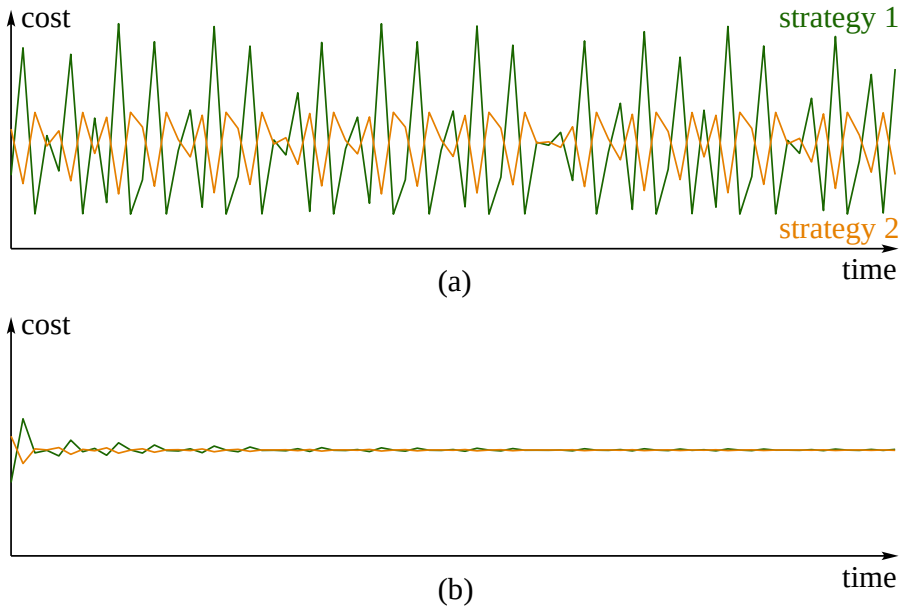

FIGURE 3. Costs and time average costs for Example 1. The parameter A is fixed at $A = 12.2$. After 10^4 burn-in iterates, for the next 75 iterates we plot costs $C(j)(T)$ $(T = 1, 2, \ldots, 75)$ of both strategies in (a) and their time averages $(1/T)\sum_{n=1}^{T} C(j)(n)$ in (b). We see that the convergence of time-average costs, predicted in Corollary 3.11, is quite fast.

case (consisting of one type of learning agents), which also coincides with the Nash equilibrium of the congestion game, see Ref. [9].

Our results provide an explicit example of learning in games such that a classic game-theoretic concept of equilibria agrees with a macroscopic order (time-average total flow and time-average cost) that arises from a plausible non-equilibrium discrete-time learning dynamics at the individual level. This connection between a static variable prediction framework and the actual time-average prediction adds to the ergodicity economics literature that has recently received increasing attention in economics and econophysics, see Ref. [27].

Acknowledgemets Research of Michał Misiurewicz was partially supported by grant number 426602 from the Simons Foundation. Jakub Bielawski and Fryderyk Falniowski acknowledge support from a subsidy granted to the Cracow University of Economics - Project no. 082/EIM/2022/POT. Thiparat Chotibut was supported by Thailand Science Research and Innovation Fund Chulalongkorn University (IND66230005).

REFERENCES

[1] Gabriel P Andrade, Rafael Frongillo, and Georgios Piliouras. Learning in matrix games can be arbitrarily complex. In *Conference on Learning Theory*, pages 159–185. PMLR, 2021.

[2] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

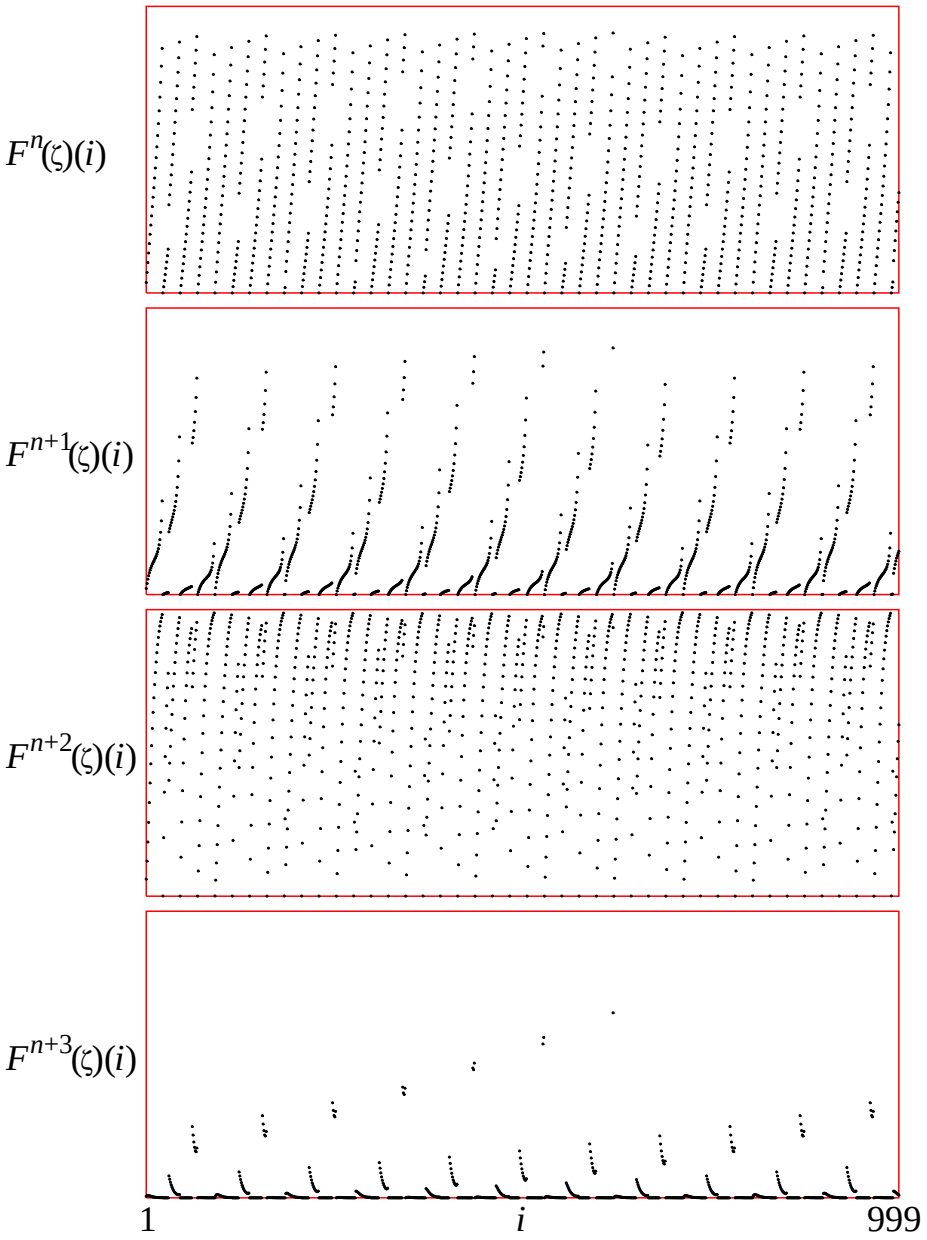

FIGURE 4. Spatiotemporal chaos can emerge from the dynamics of 999 types of agents from Example 2. The diagrams show the values of $F^{n+k}(\zeta)$ for $k = 0, 1, 2, 3$. The first 20000 iterates are the burn-in period and the plots above are the first four time steps after the burn-in; $n = 20001$.

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.

[4] Jakub Bielawski, Thiparat Chotibut, Fryderyk Falniowski, Grzegorz Kosiorowski, Michał Misiurewicz, and Georgios Piliouras. Follow-the-regularized-leader routes to chaos in routing games. In *International Conference on Machine Learning*, pages 925–935. PMLR, 2021.

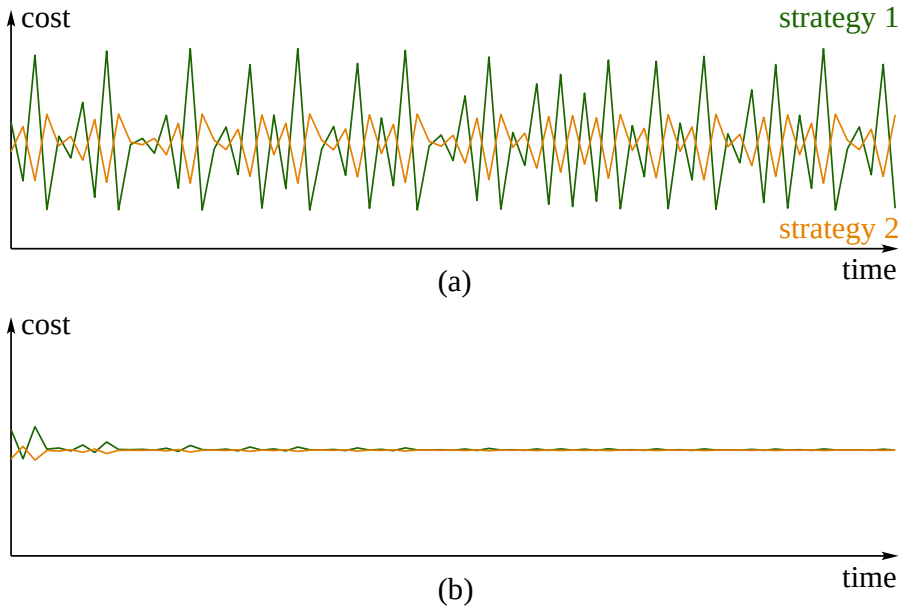

FIGURE 5. Costs and time average costs for Example 2. After 10^4 burn-in iterates, for the next 75 iterates we plot costs $C(j)(T)$ $(T = 1, 2, \ldots, 75)$ of both strategies in (a) and their time averages $(1/T)\sum_{n=1}^{T} C(j)(n)$ in (b). We see that the convergence of time-average costs, predicted in Corollary 3.11, is quite fast. One can see that the behavior observed here is very similar to the one showed in Figure 3 for Example 1.

[5] C. Camerer and T. Hua Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67:827–874, 1999.
[6] Nicolò Cesa-Bianchi and Gábor Lugoisi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
[7] Erick Chastain, Adi Livnat, Christos Papadimitriou, and Umesh Vazirani. Algorithms, games, and evolution. *Proceedings of the National Academy of Sciences*, 111(29):10620–10623, 2014.
[8] Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory*, pages 807–834. PMLR, 2019.
[9] Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. The route to chaos in routing games: When is price of anarchy too optimistic? *Advances in Neural Information Processing Systems*, 33:766–777, 2020.
[10] Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. Family of chaotic maps from game theory. *Dynamical Systems*, 36(1):48–63, 2021.
[11] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, page 133, 1999.
[12] Drew Fudenberg and David K Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7):1065–1089, 1995.
[13] Tobias Galla and J. Doyne Farmer. Complex dynamics in learning complicated games. 110(4):1232–1236, 2013.
[14] James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3(2):97–139, 1957.
[15] T. H. Ho, C. F. Camerer and J. K. Chong. Self-tuning experience weighted attraction learning in games. *J. Economic Theory*, 133:177–198, 2007.

[16] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998.
[17] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.
[18] Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 533–542, 2009.
[19] Walid Krichene, Benjamin Drighes, and Alexandre Bayen. On the convergence of no-regret learning in selfish routing. In *International Conference on Machine Learning*, pages 163–171. PMLR, 2014.
[20] R. Lahkar and R. M. Seymour. Reinforcement learning in population games. *Games Econ. Behav.*, 80:10–38, 2013.
[21] Tien-Yien Li and James A. Yorke. Period three implies chaos. *The American Mathematical Monthly*, 82(10):985–992, 1975.
[22] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, February 1994.
[23] V. Losert and E. Akin. Dynamics of games and genes: Discrete versus continuous time. *Journal of Mathematical Biology*, 1983.
[24] Archan Mukhopadhyay and Sagar Chakraborty. Deciphering chaos in evolutionary games. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(12):121104, 2020.
[25] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007.
[26] Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. In *Advances in Neural Information Processing Systems*, pages 5872–5882, 2017.
[27] Ole Peters. The ergodicity problem in economics. *Nature Physics*, 15(12):11216–1221, 2019.
[28] Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.
[29] Tim Roughgarden. *Twenty lectures on algorithmic game theory*. Cambridge University Press, 2016.
[30] Sylvie Ruette. *Chaos on the interval*, volume 67 of *University Lecture Series*. American Mathematical Society, 2017.
[31] James B. T. Sanders, J. Doyne Farmer, and Tobias Galla. The prevalence of chaotic dynamics in games with many players. *Scientific Reports*, 8, 2018.
[32] William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, 2010.
[33] Yuzuru Sato, Eizo Akiyama, and J. Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.
[34] Peter Schuster and Karl Sigmund. Replicator dynamics. *Journal of Theoretical Biology*, 100(3):533 – 538, 1983.
[35] Brian Skyrms. Chaos and the explanatory significance of equilibrium: Strange attractors in evolutionary game dynamics. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1992, pages 374–394. Philosophy of Science Association, 1992.
[36] Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(12):145 – 156, 1978.
[37] Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.
[38] J. W. Weibull. *Evolutionary Game Theory*. MIT Press; Cambridge, MA: Cambridge University Press., 1995.
[39] Y. Yang and J, Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint*, arXiv:2011.00583, 2020.

(J. Bielawski) Department of Mathematics, Cracow University of Economics, Rakowicka 27, 31-510 Kraków, Poland
E-mail address: jakub.bielawski@uek.krakow.pl

(T. Chotibut) Chula Intelligent and Complex Systems, Department of Physics, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand.
E-mail address: Thiparat.C@chula.ac.th, thiparatc@gmail.com

(F. Falniowski) Department of Mathematics, Cracow University of Economics, Rakowicka 27, 31-510 Kraków, Poland
E-mail address: falniowf@uek.krakow.pl

(M. Misiurewicz) Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, 402 N. Blackford Street, Indianapolis, IN 46202, USA
E-mail address: mmisiure@math.iupui.edu

(G. Piliouras) DeepMind, 14-18 Handyside Street, London N1C 4DN, UK
E-mail address: gpil@deepmind.com