EVOLUTION OF REWARDS IN GAMES

1. Introduction

This is an introduction to the paper.

2. Model

2.1. **Game, payoffs, and utilities.**

We consider a general N-player game where each player i has an action set $\mathcal{A}_i$ from which it can choose an action $a_i \in \mathcal{A}_i$. We think of these N players as belonging to a larger population, where a (random) matching has occurred between population members that generated many groups of N players (see below the evolutionary setting). The action space $\mathcal{A}_i$ can be a finite set of n discrete actions or a continuous set, such as a subset of $\mathbb{R}$. The environment as well as the actions of the other players affect the material payoff of every individual. There is a set Θ of environmental states, with typical element $\theta \in \Theta$ (the environment can also be continuous or discrete). The material payoff of a player is given by the function $\pi_i : \prod_{j \in N} \mathcal{A}_j \times \Theta \to \mathbb{R}$, which ultimately affects the fitness of i (see below for details). To simplify notation, we will also write $\mathcal{A} = \prod_{j \in N} \mathcal{A}_j$ for the set of all possible action profiles.

The players have subjective preference or utility functions, $u_i : \prod_{j \in N} \mathcal{A}_j \times \Theta \to \mathbb{R}$, that may differ from the objective material payoff π_i. The utility function u_i is genetically determined and is the evolving trait: we are interested in the function(s) u_i that is(are) favored by natural selection. In the forthcoming analysis of the model, it may be easier to think of the game as having a single-valued "outcome", o that is a function of the action profile and the environment, i.e. $o(\mathbf{a}, \theta) \in O$, where the outcome space, O, is a subset of $\mathbb{R}$. The utility function may then be defined directly over outcomes $u_i : O \to \mathbb{R}$.

The game is repeated at discrete times $t = 1, 2, \ldots$ and the players choose an action profile $\mathbf{a}_t$. The environment fluctuates independently from the players' actions and takes value θ_t at time t. The probability that state θ occurs at time t is written $\rho(\theta)$, and is thus independent of time (i.i.d. environment). The players cannot observe the state of the environment prior to choosing their actions. The material payoff to player i when the action profile is $\mathbf{a}_t$ and the environment is in state θ_t is given by $\pi_i(\mathbf{a}_t, \theta_t)$, which we may also write $\pi_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$, where $\mathbf{a}_{-i,t} \in \prod_{\substack{j \in N \\ j \neq i}} \mathcal{A}_j$ is the action profile of all players except player i. Each individual observes privately the utility $u_i(\mathbf{a}_t, \theta_t)$ which results from the game outcome $o_t = (\mathbf{a}_t, \theta_t)$ at time t.

2.2. **Learning.**

The learning model is taken from Dridi & Lehmann (2014), where players use the material payoffs of the game to update motivations about actions. Our learning model will not be very different from this previous work because all we have to change is that individuals update their choice probabilities using the subjective utility of a game outcome, rather than the objective material payoff. Hence, the difference between this previous paper and our formulation of learning dynamics is that in Dridi & Lehmann (2014), the games considered are symmetric, while here the utility function of each player is different, which translates into an asymmetric game defined by the family of utility functions $(u_i)_{i \in N}$.

Date: August 9, 2016.

While the general model of Dridi & Lehmann (2014) allows for both trial-and-error learning and belief-based learning, we will first only consider trial-and-error learning. At every time t of the repeated game defined above, each individual i holds in memory action values $V_{i,t}(a_i)$ for all actions $a_i \in \mathcal{A}_i$. The learning rule of individual i is to update action values according to

$$V_{i,t+1}(a_i) = V_{i,t}(a_i) + \gamma_t \mathbb{1}(a_i, a_{i,t}) u_i(a_i, \mathbf{a}_{-i,t}, \theta_t), \tag{1}$$

where $\gamma_t \in (0,1)$ is a decreasing learning rate in the sense of stochastic approximation theory. This decreasing learning rate allows us to approximate the above stochastic difference equation with a deterministic mean-field differential equation that asymptotically tracks the original stochastic dynamics. The expression $\mathbb{1}(a_i, a_{i,t})$ is an indicator variable that equals 1 if $a_i = a_{i,t}$, and 0 otherwise. While eq. **??** is not very often used in the literature, it can be seen as a conscious updating of action values because an action that is not played at time t keeps the same action value at time $t+1$. Traditionally in reinforcement learning models, actions that are not played for a long period of time tend to be forgotten, and hence their values come back to 0. Our model captures more learning processes where early experience is critical in determining stable outcomes, and hence may be used to capture fast learning dynamics rather than lifelong learning processes.

We generally call the right-hand side of eq. **??** the learning rule, written $\ell_i(V_i, h_i)$, of individual i. The learning rule takes the previous vector of action values V_i and a new information h_i (in our present model, $h_i = u_i(\cdot, \mathbf{a}_{-i,t}, \theta_t)$, but one could think of more general updating rules) and outputs a new vector of action values. Player i chooses an action $a_{i,t}$ at time t with a probability that depends on its action values $V_{i,t} = \{V_{i,t}(a_i)\}_{a_i \in \mathcal{A}_i}$. One possibility is to assume a perturbed maximization scheme which gives rise to the logit-choice function,

$$p_{i,t}(a_i) = \frac{\exp[\xi V_{i,t}(a_i)]}{\sum_{b_i \in \mathcal{A}_i} \exp[\xi V_{i,t}(b_i)]}, \tag{2}$$

where ξ is the exploration parameter (the inverse $1/\xi$ can be seen as the noise level) in choosing actions. We write $p_{i,t}$ without the action argument to denote the whole vector of action probabilities of player i.

3. Fecundity

We define the total payoff of individual i as the average material payoff obtained at equilibrium of the learning process, i.e.

$$f_i = f(u_i) = \int_{\mathcal{A}} \sum_{\theta \in \Theta} \hat{\mathbf{p}}(\mathbf{a}) \rho(\theta) \pi_i(\mathbf{a}, \theta) \, \mathrm{d}\mathbf{a}, \tag{3}$$

where $\hat{\mathbf{p}}(\mathbf{a}) = \prod_{j \in N} \hat{p}_j(a_j)$ is the equilibrium probability of action profile $\mathbf{a} = (a_1, \ldots, a_N)$. This is the product of individuals' equilibrium action probabilities $\hat{p}_j(a_j)$. (In the continuous action setting, we always assume that the probability distribution over actions admits a density). Importantly, while the utility function does not appear on the rhs of eq. **??**, we still defined it as $f(u_i)$ because the equilibrium choice probabilities of a player, $\hat{p}_i(a_i)$, implicitly depend on the utility function of player i.

4. Preference evolution and the set of possible utility functions

Consider a very large population and assume that groups of N players are randomly formed at every generation to play the above repeated game. An individual's genotype corresponds to its utility function $u_i(\cdot)$, which he transmits to its offspring. The number of offspring he produces depends on its material payoffs obtained during the game as defined in eq. **??**. In the most general case, the set in which evolution occurs is the set of all possible utility functions $u : \prod_{j \in N} \mathcal{A}_j \times \Theta \to \mathbb{R}$. Let U denote such a set. When the action space is discrete, the space of utility functions is $\mathbb{R}^{|O|}$, where $O = \prod_{j\in N} \mathcal{A}_j$ is the set of possible game outcomes. Hence the dimension of the state space is equal to $|O| = \prod_{j\in N} |\mathcal{A}_j|$. In a symmetric game where all N players have n actions, the dimension of U is thus equal to $|O| = n^N$.

5. Symmetric two-player games

In this section we consider a reduction of the model for symmetric constant two-player games. The action set for each player i is $\mathcal{A}_i = \{1, 2\}$, where we identify action 1 with cooperation, and action 2 with defection.

5.1. **Four-dimensional adaptive dynamics.** The set U consists of all possible 2×2 real matrices, which can be identified with the set $\mathbb{R}^{2\times 2} = \mathbb{R}^4$. The genotype of an individual can thus also be seen as the vector $\mathbf{u} = (u_{11}, u_{12}, u_{21}, u_{22})$, where u_{ij} denotes the utility to the focal individual when he chooses action i and his opponent chooses action j.

With such definitions, one can use standard approaches in evolutionary biology, such as adaptive dynamics, in order to find the ESS utility matrix, and the analysis can be complemented to look at convergence stability. Under this approach one considers a resident population with utility $\mathbf{u}$ and asks whether a mutant with utility $\mathbf{u} + \boldsymbol{\delta}$ can invade or not. One advantage of this approach in our learning context is that the "utility game" between the resident and the mutant is quasi-symmetric, which implies that the learning dynamics determining behavior is a family of dynamics for quasi-symmetric games. This is a substantial simplification of the analysis. Save the bifurcations, the quasi-symmetric game between $\mathbf{u}$ and $\mathbf{u} + \boldsymbol{\delta}$ is very close to the symmetric game defined by $\mathbf{u}$, because the vector field determining the learning dynamics is continuous in $\mathbf{u}$.

A candidate ESS is such that

$$\left.\frac{\partial f}{\partial \mathbf{u}}\right|_{\mathbf{u}=\mathbf{u}^*} = 0. \tag{4}$$

Even before embarking on the analysis of eqs. **??**–**??**, one can remark that it is not certain that a unique strategy $\mathbf{u}^*$ can be found using standard maximization approaches. Indeed, there are discontinuities in f occurring when at least one utility changes sign. This is because f depends on the dynamical system describing the behavioral interactions between learners in the population. The behavioral dynamics actually changes completely (i.e. stable equilibria appear/disappear) when a utility changes sign, which means that there are bifurcations at the 0 axes (when $u_{ij} = 0$). In order to perform a full analysis of the model, one has to deal with the bifurcations occurring when at least when utility changes sign. These complications are best illustrated in the simplified model that we study below.

5.2. Analysis of the behavioral interactions between the strategies.

5.2.1. *Generic analysis of a 2-player interaction.* In order to compute fitness for a player, one needs to predict the outcome of behavioral interactions when this player is matched with another player that has a different utility function. The behavioral interaction between a reinforcement learner with utilities $\mathbf{u} = (u_{11}, u_{12}, u_{21}, u_{22})$ and another reinforcement learner with utilities $\mathbf{v} = (v_{11}, v_{12}, v_{21}, v_{22})$ determines a dynamical system in $[0, 1]^2$. By a slight abuse of notation, let denote these two players u and v and their probabilities to cooperate by p_u and p_v respectively. The utilities $\mathbf{u}$ and $\mathbf{v}$ themselves define an asymmetric 2-player game:

$$\begin{bmatrix} (u_{11}, v_{11}) & (u_{12}, v_{21}) \\ (u_{21}, v_{12}) & (u_{22}, v_{22}) \end{bmatrix}, \tag{5}$$

where player u chooses a row and player v chooses a column. Remark that the notation for utilities are player-centered: x_{ij} is the utility to player x when he plays i and his opponent plays j. However, when we refer to an outcome of the game without explicitly mentioning the players, we write it in the form (a_u, a_v), treating player u as the default focal player. Analyzing a generic behavioral interaction between two reinforcement learners with arbitrary utilities amounts to analyzing the behavioral dynamics of reinforcement learning in arbitrary two-player asymmetric games.

Using stochastic approximation theory, one obtains a system of differential equations describing the long-run dynamics of eqs. **??**–**??**, which reads

$$\begin{aligned}\dot{p}_u &= p_u(1-p_u)\xi\left[p_u\{p_v u_{11} + (1-p_v)u_{12}\} - (1-p_u)\{p_v u_{21} + (1-p_v)u_{22}\}\right],\\ \dot{p}_v &= p_v(1-p_v)\xi\left[p_v\{p_u v_{11} + (1-p_u)v_{12}\} - (1-p_v)\{p_u v_{21} + (1-p_u)v_{22}\}\right].\end{aligned} \tag{6}$$

In Fig. **??**, we show the ten possible equilibria of eq. **??**. We call these the generic behavioral equilibria. Depending on the specific values of **u** and **v**, some of these equilibria may not exist anymore.

The sign of the utilities **u** and **v** play a fundamental role in determining the stability of the different behavioral equilibria. Indeed, one has that a pure equilibrium is locally stable if and only if both players have a positive utility for this outcome. In other words, if players u and v do not "agree" on preferred outcomes, then a pure behavioral equilibrium cannot be stable. This intuitive result is mathematically true because the eigenvalues of the Jacobian matrix evaluated at a pure outcome (i, j) are simply

$$\lambda_1 = -\xi u_{ij}, \lambda_2 = -\xi v_{ji}. \tag{7}$$

This has important evolutionary consequences. Under the assumption that fitness is evaluated only at equilibrium of the learning process, it means that as long as two strategies have utilities of the same signs, none has an evolutionary advantage over the other (except for strategies in the "Mismatcher" class, $\text{sign}(\mathbf{u}) = (-,+,-,+)$, or "Reluctant" strategies, $\text{sign}(\mathbf{u}) = (-,-,-,-)$). Mixed behavioral equilibria (i.e., equilibria where at least one player has a mixed strategy) require a little more work. First one notices that the four mixed equilibria on the boundaries exist only if the utilities of the mixing player have the same sign. For instance, the equilibrium $\left(\frac{u_{21}}{u_{11}+u_{21}}, 1\right)$ on the top boundary of the state space exists only when u_{11} and u_{21} have the same sign.

5.2.2. *Subset of strategies.* In this section we calculate the payoffs and fitnesses for all possible behavioral interactions between the 4 strategies Realistic, Other-regard, Selfish, and Manipulator. From these calculations, we derive invasion conditions within this set of strategies.

Some assumptions need to be made in order to compute analytical expressions for long-term payoffs. First, when the behavioral dynamics admit several stable equilibria, typically the stochastic dynamics may converge to any of these stable equilibria. It may not always be possible to know which particular equilibrium will be reached by the stochastic process, because the trajectory may go out from the basin of attraction of a stable equilibrium under the influence of large stochastic shocks (these large shocks are more likely to occur at the beginning of a behavioral interaction). For these reasons, we assume that the initial preferences of players are unbiased, such that the process starts in the center of the state space ($p_1 = p_2 = 1/2$). Also, we assume that the process may reach each equilibrium with equal probability; in other words, the distribution of stable equilibria is uniform. Such an assumption is valid for relatively high values of the exploration rate, $\xi \approx 10$ (Dridi & Lehmann, 2015).

6. Continuous action space

6.1. **Normally distributed actions.** Here, we treat the case where the action space is one-dimensional continuous, e.g., when $\mathcal{A}_i \subseteq \mathbb{R}$ for all i. This allows to capture investment games such as a continuous public goods game. We adapt a model of Beigy & Meibody (2006), who study learning of a single decision-maker who faces a stochastic stationary environment. In their setting, the decision-maker tries to minimize the expectation of a given cost function where at each discrete time step he can exert an action $a \in \mathbb{R}$. In order to adapt their model to multi-player games, only a few tweaks need to be made.

We consider that at each time step t, every player is characterized by a mean action $m_{i,t}$, and a standard deviation $\sigma_{i,t}$. The action $a_{i,t}$ chosen by i at time t is then drawn from a normal distribution with mean $m_{i,t}$ and standard deviation $\sigma_{i,t}$, i.e., $a_{i,t} \sim \mathcal{N}(m_{i,t}, \sigma_{i,t})$. A possible interpretation of this model of action choice is that $m_{i,t}$ represents the preferred action by player i, which he would like to implement but there is a stochastic perturbation during action choice, such that $a_{i,t} = m_{i,t} + \varepsilon_{i,t}$, where $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_{i,t})$ is independent across players and time steps. With this interpretation, $\varepsilon_{i,t}$ can be seen as an exogenous noise during action implementation, or can also be seen as explicit exploration of the player around its preferred action (note that this model of action perturbation is different from the above model of payoff perturbation that generates the logit choice rule, eq. **??**).

Since a normal distribution is fully specified by its mean and standard deviation, in order to know the action distribution of a player at time $t + 1$, we only need to know how $m_{i,t+1}$ and $\sigma_{i,t+1}$ are updated given the previous values at time t and the outcome of the game. We will use the following updating rule for the mean action,

$$m_{i,t+1} = m_{i,t} + \gamma_t u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t) \left(\frac{a_{i,t} - m_{i,t}}{\sigma_{i,t}} \right), \tag{8}$$

which essentially entails that the individual wants his new mean action $m_{i,t+1}$ to come closer to the chosen action $a_{i,t}$ if the obtained utility $u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$ is positive, while he wants his mean action to move away from $a_{i,t}$ if the obtained utility is negative; the magnitude of such movement is proportional to the actual magnitude of the obtained utility. Consequently, this is one of the simplest way of capturing in a continuous action setting the biological notion of approach towards rewards and avoidance of punishments. The standard deviation $\sigma_{i,t}$ appears in the denominator of the reinforcement in eq. **??** so that the individual accounts for its own exploration strategy when updating action value: if $\sigma_{i,t}$ is large, then there is a large probability that $a_{i,t}$ will be far from the mean, so the utility obtained by the individual should not be interpreted as reflecting a wrong mean action $m_{i,t}$ (the utility is mainly a reflection of the great variance in action choice), hence the update should be minor. Finally, the learning rate γ_t must obey the assumptions of stochastic approximation theory, just as in the discrete-action model (eq. **??**) in order to ensure convergence of the learning process.

In the case where $\sigma_{i,t} = \sigma_i$ is a constant, independent of time, then σ_i acts as an additional learning rate, that scales the speed at which the mean-field deterministic dynamics converge to an equilibrium: larger variance should lead to faster convergence (see the analysis below for details).

A technical assumption needed to make the analysis tractable is that $u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$ is twice continuously differentiable in the actions of all the players (note that this may create some problems when defining games with bounded action space, where typically the payoff function is not differentiable on the boundaries).

The vector $\boldsymbol{m}_t = (m_{1,t}, \ldots, m_{N,t})$ collects the mean action of all players at time t.

6.2. **Stochastic approximation.** Let $R_{i,t+1} = u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)\left(\frac{a_{i,t}-m_{i,t}}{\sigma_{i,t}}\right)$ be the reinforcement to the mean action of player i, and let $\mathbf{R}_{t+1} = (R_{i,t+1})_{i\in N}$ denote the vector collecting the reinforcements of all the players at time $t+1$. Stochastic approximation theory tells us that $\mathbb{E}[\mathbf{R}_{t+1}|\boldsymbol{m}_t]$ is a vector field that defines a differential equation for $\boldsymbol{m}$ whose asymptotic path is followed by the original stochastic process $\{\boldsymbol{m}_t\}_{t\geq 0}$. To obtain an explicit expression for $\mathbb{E}[\mathbf{R}_{t+1}|\boldsymbol{m}_t]$, we use a second-order Taylor series expansion of $u_i(a_{i,t}, \mathbf{a}_{-i,t}, \theta_t)$ around $m_{i,t}$, which gives

$$\mathbb{E}[R_{i,t+1}|\boldsymbol{m}_t] = \mathbb{E}\Big[\Big(u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t) + \nabla u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)^{\mathrm{T}}(\mathbf{a}_t - \boldsymbol{m}_t) \\ + (\mathbf{a}_t - \boldsymbol{m}_t)^{\mathrm{T}}\mathbf{H}u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{m}_t)\Big)\frac{(a_{i,t} - m_{i,t})}{\sigma_i}\Big|\boldsymbol{m}_t\Big], \quad (9)$$

where $\nabla u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t) = \left(\frac{\partial u_i}{\partial m_{1,t}}, \ldots, \frac{\partial u_i}{\partial m_{N,t}}\right)$ is the gradient vector and $\mathbf{H}u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)$ is the Hessian matrix of u_i, with typical element $[\mathbf{H}u_i]_{kj} = \frac{\partial^2 u_i}{\partial m_{k,t}\partial m_{j,t}}$. The vector $\mathbf{a}_t - \boldsymbol{m}_t$ has typical element $a_{j,t} - m_{j,t}$. Taking the non-random terms out of the expectation, the above equation can be rewritten as

$$\mathbb{E}[R_{i,t+1}|\boldsymbol{m}_t] = \frac{1}{\sigma_i}u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)\mathbb{E}\Big[(a_{i,t} - m_{i,t})\Big|\boldsymbol{m}_t\Big] + \nabla u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)\mathbb{E}\Big[(a_{i,t} - m_{i,t})(\mathbf{a}_t - \boldsymbol{m}_t)\Big|\boldsymbol{m}_t\Big] \\ + \mathbb{E}\Big[(a_{i,t} - m_{i,t})(\mathbf{a}_t - \boldsymbol{m}_t)^{\mathrm{T}}\mathbf{H}u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{m}_t)\Big|\boldsymbol{m}_t\Big]. \quad (10)$$

In order to find expressions for the various expectations above it will be useful to keep in mind that for each player i, the random variable $a_{i,t} - m_{i,t} = \varepsilon_{i,t}$ is normally distributed with zero mean and variance σ_i^2. In particular, this implies that $\mathbb{E}\Big[(a_{i,t} - m_{i,t})\Big|\boldsymbol{m}_t\Big] = 0$ for all i, so that $\mathbb{E}\Big[(\mathbf{a}_t - \boldsymbol{m}_t)\Big|\boldsymbol{m}_t\Big] = (0, 0, \ldots, 0)$. Also

$$\begin{aligned}\mathbb{E}\Big[(a_{i,t} - m_{i,t})(\mathbf{a}_t - \boldsymbol{m}_t)\Big|\boldsymbol{m}_t\Big] &= \Big(\mathrm{Cov}[a_{1,t}, a_{i,t}|\boldsymbol{m}_t], \ldots, \mathrm{Var}[a_{i,t}|\boldsymbol{m}_t], \ldots, \mathrm{Cov}[a_{N,t}, a_{i,t}|\boldsymbol{m}_t]\Big) \\ &= \Big(0, \ldots, \sigma_i^2, \ldots, 0\Big), \quad (11)\end{aligned}$$

because action choice is independent across players. (An interesting note here on social learning is that if players copy each other then $\mathrm{Cov}[a_{j,t}, a_{i,t}|\boldsymbol{m}_t]$ of players i and j may be different than 0, depending on whether they can copy one another.) The term involving the Hessian matrix has the form

$$\mathbb{E}\Big[(a_{i,t} - m_{i,t})(\mathbf{a}_t - \boldsymbol{m}_t)^{\mathrm{T}}\mathbf{H}u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{m}_t)\Big|\boldsymbol{m}_t\Big] = \\ \sum_{j\in N}\sum_{k\in N}\frac{\partial^2 u_i}{\partial m_{k,t}\partial m_{j,t}}\mathbb{E}\Big[(a_{i,t} - m_{i,t})(a_{j,t} - m_{j,t})(a_{k,t} - m_{k,t})\Big|\boldsymbol{m}_t\Big], \quad (12)$$

where the triplet covariance $\mathbb{E}\Big[(a_{i,t} - m_{i,t})(a_{j,t} - m_{j,t})(a_{k,t} - m_{k,t})\Big|\boldsymbol{m}_t\Big]$ is zero again because action choice is independent across players whenever at least one of $j \neq i$ or $k \neq i$ or $k \neq j$ is true. When $j = i$ and $k = i$, this triplet equals $\mathbb{E}\Big[(a_{i,t} - m_{i,t})^3\Big|\boldsymbol{m}_t\Big] = 0$ (because the odd moments of a centered Normal distribution are 0). Thus we have

$$\mathbb{E}\Big[(a_{i,t} - m_{i,t})(\mathbf{a}_t - \boldsymbol{m}_t)^{\mathrm{T}}\mathbf{H}u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)(\mathbf{a}_t - \boldsymbol{m}_t)\Big|\boldsymbol{m}_t\Big] = 0. \quad (13)$$

Plugging back the expressions of these expectations in eq. **??**, the expected reinforcement of player i takes the simple form

$$\mathbb{E}[R_{i,t+1}|\boldsymbol{m}_t] = \sigma_i\frac{\partial u_i(m_{i,t}, \boldsymbol{m}_{-i,t}, \theta_t)}{\partial m_{i,t}}. \quad (14)$$

Thus, the differential equation

$$\frac{\mathrm{d}m_i}{\mathrm{d}t} \equiv \dot{m}_i = \sigma_i \frac{\partial u_i(m_i, \boldsymbol{m}_{-i}, \theta)}{\partial m_i}, \qquad i \in N \tag{15}$$

describes the long-run stochastic dynamics of the mean action of player i. Eq. **??** is a gradient dynamical system, where every player can be seen as maximizing his own utility function. In previous work, such gradient dynamics have been used as an *ad hoc* model of learning dynamics. It is interesting that we recover this result by explictly modeling the stochastic learning dynamics.

A counterintuitive feature of eq. **??**, when we compare it to eq. **??**, is that according to eq. **??**, learning speed should increase with the standard deviation σ_i, while in the original stochastic updating rule (eq. **??**), larger σ_i should lead to smaller updates to the mean action $m_{i,t+1}$.

It is noteworthy that at an equilibrium for the means $\hat{\boldsymbol{m}}$, action choice is still stochastic so that the equilibrium action of every player is $\hat{a}_i \sim \mathcal{N}(\hat{m}_i, \sigma_i)$.

6.3. **Bounded action space.** A difficulty that may arise from defining the chosen action as a normal random variable is that normal random variables are defined over the entire real line, $\mathbb{R}$. In certain cases, we may want to consider a bounded state space, of the form $(0, M)$. One way to achieve this is to make the transformation

$$a_i = \frac{M}{1 + \exp[-\tilde{a}_i]}, \tag{16}$$

where $\tilde{a}_i \in \mathbb{R}$ is the variable that is dynamically updated in eq. **??**, and $a_i \in (0, M)$ is the action of individual i. For instance in a public-goods game, M may represent the endowment of each player, which is the maximum contribution he can make to the public good.

6.4. **Baseline simple individual decision problem.** In this section we display simulations of the learning algorithm in the 1-player case under a deterministic environment. The utility function used here is given by

$$u_i(a_i) = -(a_i - \beta)^2, \tag{17}$$

which displays a single global optimum action achieved at $a_i = \beta$. Though there is nothing particularly exciting about this model, it is used only to show numerically that our stochastic approximation results presented above do hold and help predict the real dynamics of the players' actions.

TO DO: test in multi-player games with more complex payoff functions.

7. Tables

Table 1. Classification of behavioral outcomes in the discrete action model amongst the 4 strategies considered.

Interaction	Always stable equilibria	Sometimes stable equilibria	Stability condition
R vs. R	$(1,1)$ and $(0,0)$		
R vs. O	$(0,1)$ and $(1,1)$		
R vs. M	$(1,1)$ and $(0,\frac{v_{22}}{v_{12}+v_{22}})$		
R vs. S	$(0,0)$		
O vs. O	$(1,1)$		
O vs. M	$(1,0)$ and $(1,1)$		
O vs. S	$(1,0)$		
M vs. M	$(1,1)$	$(\frac{u_{22}}{u_{12}+u_{22}},0)$ and $(0,\frac{v_{22}}{v_{12}+v_{22}})$	$\lvert u_{12}\rvert < \lvert u_{21}\rvert$, $\lvert v_{12}\rvert < \lvert v_{21}\rvert$,
M vs. S	$(\frac{u_{22}}{u_{12}+u_{22}},0)$		
S vs. S	$(0,0)$		

Table 2. Generic total payoff in the discrete action model amongst the 4 strategies considered.

Interaction	Generic fitness
R vs. R	$(\frac{1}{2}(\mathcal{R}+\mathcal{P}),\frac{1}{2}(\mathcal{R}+\mathcal{P}))$
R vs. O	$(\frac{1}{2}(\mathcal{R}+\mathcal{T}),\frac{1}{2}(\mathcal{R}+\mathcal{S}))$
R vs. M	$\left(\frac{1}{2}\left(\mathcal{R}+\mathcal{T}\left(\frac{v_{22}}{v_{12}+v_{22}}\right)+\mathcal{P}\left(1-\frac{v_{22}}{v_{12}+v_{22}}\right)\right),\frac{1}{2}\left(\mathcal{R}+\mathcal{S}\left(\frac{v_{22}}{v_{12}+v_{22}}\right)+\mathcal{P}\left(1-\frac{v_{22}}{v_{12}+v_{22}}\right)\right)\right)$
R vs. S	$(\mathcal{P},\mathcal{P})$
O vs. O	$(\mathcal{R},\mathcal{R})$
O vs. M	$(\frac{1}{2}(\mathcal{R}+\mathcal{S}),\frac{1}{2}(\mathcal{R}+\mathcal{T}),)$
O vs. S	$(\mathcal{S},\mathcal{T})$
M vs. M	$(\mathcal{R},\mathcal{R})$ or ...
M vs. S	$\left(\mathcal{S}\left(\frac{u_{22}}{u_{12}+u_{22}}\right)+\mathcal{P}\left(1-\frac{u_{22}}{u_{12}+u_{22}}\right),\mathcal{T}\left(\frac{u_{22}}{u_{12}+u_{22}}\right)+\mathcal{P}\left(1-\frac{u_{22}}{u_{12}+u_{22}}\right)\right)$
S vs. S	$(\mathcal{P},\mathcal{P})$

Table 3. Total payoff in the 3 different types of games in the discrete action model amongst the 4 strategies considered.

Interaction	Prisoner's Dilemma	Stag Hunt	Snowdrift
R vs. R	$(\frac{b-c}{2}, \frac{b-c}{2})$	$(\frac{(k+1)(b-c)}{2k}, \frac{(k+1)(b-c)}{2k})$	$(\frac{2b-c}{4}, \frac{2b-c}{4})$
R vs. O	$(b - \frac{c}{2}, \frac{b-c}{2})$	$(\frac{(k+1)(b-c)}{2k}, \frac{b-c}{2})$	$(b - \frac{c}{4}, b - \frac{3c}{4})$
R vs. M	$(\frac{(b-c)v_{12}+(2b-c)v_{22}}{2(v_{12}+v_{22})}, \frac{(b-c)v_{12}+(b-2c)v_{22}}{2(v_{12}+v_{22})})$	$(\frac{(b-c)k+(b-c)}{2k}, \frac{(b-c+(b-c)k)v_{12}+(b-2c)kv_{22}}{2k(v_{12}+v_{22})})$	$(\frac{(2b-c)v_{12}+(4b-c)v_{22}}{4(v_{12}+v_{22})}, \frac{(2b-c)v_{12}+(4b-3c)v_{22}}{4(v_{12}+v_{22})})$
R vs. S	$(0, 0)$	$(\frac{b-c}{k}, \frac{b-c}{k})$	$(0, 0)$
O vs. O	$(b - c, b - c)$	$(b - c, b - c)$	$(b - \frac{c}{2}, b - \frac{c}{2})$
O vs. M	$(\frac{b}{2} - c, b - \frac{c}{2})$	$(\frac{b}{2} - c, \frac{(b-c)k+b-c}{2k})$	$(b - \frac{3c}{4}, b - \frac{c}{4})$
O vs. S	$(-c, b)$	$(-c, \frac{b-c}{k})$	$(b - c, b)$
M vs. M	$(b - c, b - c)$	$(b - c, b - c)$	$(b - \frac{c}{2}, b - \frac{c}{2})$
M vs. S	$(\frac{-cu_{22}}{u_{12}+u_{22}}, \frac{bu_{22}}{u_{12}+u_{22}})$	$(\frac{(b-c)u_{12}-cku_{22}}{k(u_{12}+u_{22})}, \frac{b-c}{k})$	$(\frac{(b-c)u_{22}}{u_{12}+u_{22}}, \frac{bu_{22}}{u_{12}+u_{22}})$
S vs. S	$(0, 0)$	$(\frac{b-c}{k}, \frac{b-c}{k})$	$(0, 0)$

8. Figures

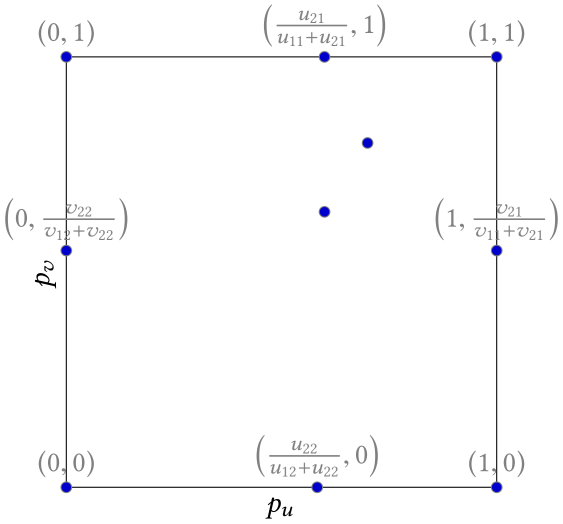
$(0, 1)$
$\left(\frac{u_{21}}{u_{11}+u_{21}}, 1\right)$
$(1, 1)$
$\left(0, \frac{v_{22}}{v_{12}+v_{22}}\right)$
$\left(1, \frac{v_{21}}{v_{11}+v_{21}}\right)$
p_v
$(0, 0)$
$\left(\frac{u_{22}}{u_{12}+u_{22}}, 0\right)$
$(1, 0)$
p_u

Figure 1. The ten generic behavioral equilibria in a 2×2 game. The two interior equilibria have long expressions that are not shown here.

C, C	C, D
D, C	D, D

Outcome matrix

$b - c$	$-c$
$\frac{b-c}{k}$	$\frac{b-c}{k}$

Stag hunt

$b - c$	$-c$
b	0

Prisoner's Dilemma

$b - \frac{c}{2}$	$b - c$
b	0

Snowdrift game

$b > c > 0, k > 1$

+	−
+	0

Realistic

+	+
−	−

Other-regard

−	−
+	+

Selfish

+	−
+	−

Manipulator

Figure 2. The possible outcomes, the 3 different fitness games, and the 4 strategies considered in the discrete action model. A strategy is defined by the outcomes to which it associates a positive or negative utility in the corresponding outcome matrix (top). Cooperation is denoted by C and defection by D. In the outcome matrix, the first letter refers to the action of the focal player (row) and the second letter to the action of its opponent (column). For example, the strategy Realistic associates a positive utility to the outcomes that yield positive material payoffs, and negative utility to outcomes yielding negative material payoffs. The strategy Other-regard associates a positive utility to the outcome where both the focal player cooperates $(C, \cdot)$ and has a negative utility for outcomes where the focal player defects $(D, \cdot)$.

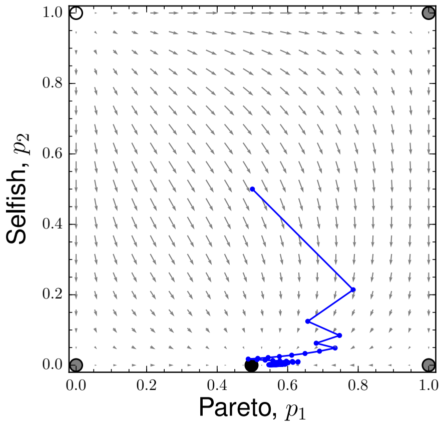

Figure 3. Vector field (gray arrows) and stochastic trajectory (blue line) for the interaction between "Pareto" and "Selfish". On the x-axis is represented the probability that Pareto cooperates (p_1), while on the y-axis, this is the probability that Selfish cooperates (p_2). The stochastic trajectory is started from the center of the state space $(p_1, p_2) = (\frac{1}{2}, \frac{1}{2})$ and dots on it represent interaction rounds between the players. Circles represent equilibria: a white-filled circle is a source (both associated eigenvalues are positive); a gray-filled circle is a saddle (one positive and one negative associated eigenvalue); a black circle is a sink (both associated eigenvalues are negative).

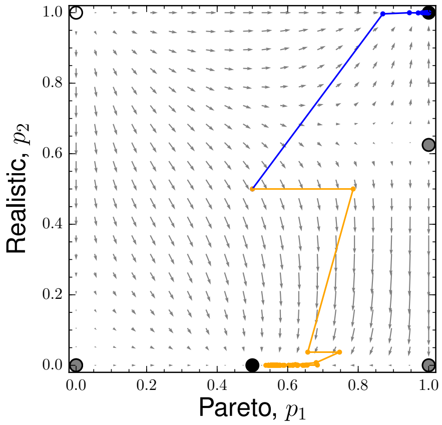

Figure 4. Same as Fig. **??** but for the interaction between Pareto and Realistic. Here the deterministic mean field equation admits two locally stable equilibria. The two stochastic trajectories of different color correspond to simulation runs that respectively converge to one of the locally stable equilibria.

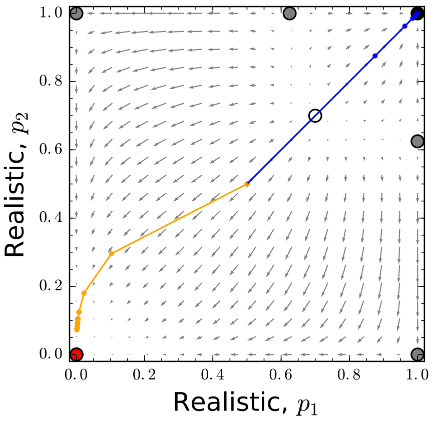

Figure 5. Same as Fig. **??** but for the interaction between Realistic and Realistic. The red-filled dot denotes an equilibrium with 0 eigenvalues.

dyn4strat.pdf

Figure 6. Vector field for the replicator dynamics in the 4-strategy game defined by the competition between Realistic, Other-regard, Manipulator, and Selfish, when the underlying one-shot fitness game is a Prisoner’s dilemma.

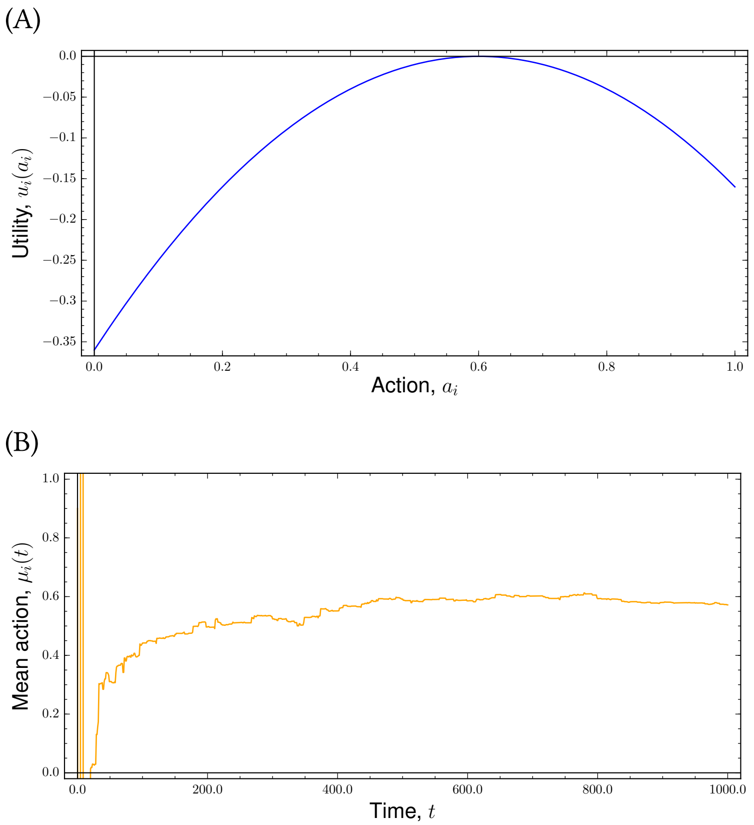

FIGURE 7. Continuous learning in a deterministic 1-player game. (A) The function $u_i(a_i)$ to be maximized by the player (eq. ??), with $\beta = 0.6$. (B) Typical dynamics of the mean action, $m_{i,t+1}$ when the utility function is that in (A).