Supplementary Information

Introspection dynamics in asymmetric multiplayers games

Marta Couto[1] and Saptarshi Pal[1]

[1]Max Planck Research Group Dynamics of Social Behavior, Max Planck Institute for Evolutionary Biology, 24306 Ploen, Germany

Abstract

Introduction

Model

We consider the normal form game with N players where $N > 2$. In the game, a player, say player i, can play actions from their action set, $\mathbf{A}_i := \{a_{i,1}, a_{i,2}, ..., a_{i,m_i}\}$. The action set of player i has m_i actions. There are, therefore, $m_i \times m_2 \times ... \times m_N$ distinct states of the game. We denote a state of the game by collecting the actions of all the players in the game in a vector, $\mathbf{a} := (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N)$ where $\mathbf{a} \in \mathbf{A} := \mathbf{A}_1 \times \mathbf{A}_2 \times ... \times \mathbf{A}_N$. We also use the notation, $\mathbf{a} := (\mathbf{a}_i, \mathbf{a}_{-i})$ to denote the state from the perspective of player i. In the state $(\mathbf{a}_i, \mathbf{a}_{-i})$, player i plays the action $\mathbf{a}_i \in \mathbf{A}_i$ and their co-players play the action $\mathbf{a}_{-i} \in \mathbf{A}_{-i} := \prod_{j \neq i} \mathbf{A}_j$. The payoff of a player depends on the state of the game. We denote the payoff of player i in the state $\mathbf{a}$ with $\pi_i(\mathbf{a})$ or $\pi_i(\mathbf{a}_i, \mathbf{a}_{-i})$.

The players update their strategies over time using the introspection dynamics[1]. At every time step, one randomly chosen player can update their strategy. The randomly chosen player, say i, currently playing strategy $a_{i,k}$, compares their current payoff to the payoff that they would obtain if they played a randomly selected strategy, $a_{i,l} \neq a_{i,k}$, from their action set $\mathbf{A}_i$. This comparison is done while assuming that the co-players do not change their respective plays. When the co-players of player i play $\mathbf{a}_{-i}$, player i changes to the new strategy $a_{i,l}$ from $a_{i,k}$ with the probability,

$$p_{a_{i,k} \to a_{i,l}}(\mathbf{a}_{-i}) = \frac{1}{1 + e^{-\beta(\pi_i(a_{i,l}, \mathbf{a}_{-i}) - \pi_i(a_{i,k}, \mathbf{a}_{-i}))}} \tag{1}$$

in the next round. Here $\beta \in [0, \infty)$ is the selection strength parameter that represents the importance that players give to payoff differences while updating their strategies. At $\beta = 0$, players update to a randomly chosen strategy with probablity 0.5. For $\beta > 0$, players update to new strategy with probablity greater than 0.5 (or less than 0.5) if the switch gives them a non-zero increase (or decrease) in the payoffs. The probability that the chosen player makes no-update is then,

$$p_{a_{i,k} \to a_{i,k}}(\mathbf{a}_{-i}) = 1 - \sum_{k \neq l} p_{a_{i,k} \to a_{i,l}}(\mathbf{a}_{-i}) \tag{2}$$

Introspection dynamics can be studied by analyzing properties of the resulting transition matrix, $\mathbf{T}$. The transition matrix element $\mathbf{T}_{\mathbf{a},\mathbf{b}}$ denotes the conditional probability that the game goes to the state $\mathbf{b}$ in the next round if it is in state $\mathbf{a}$ in the current round. We define the neighbourhood set of $\mathbf{a}$ as the following:

Definition 1 (Neighbourhood set of a state)**.** *The neighbourhood set of state* $\mathbf{a}$*,* $\mathrm{Neb}(\mathbf{a})$*, is defined as the following:*

$$\mathrm{Neb}(\mathbf{a}) := \{\mathbf{b} \in \mathbf{A} \setminus \{\mathbf{a}\} : (\exists j)[\mathbf{b}_{-j} = \mathbf{a}_{-j} \wedge \mathbf{b}_j \neq \mathbf{a}_j]\} \tag{3}$$

In other words, a state in $\mathrm{Neb}(\mathbf{a})$ is a state that has exactly one player playing a different action than in state $\mathbf{a}$. Consider the game where there are three players and each player has the identical action set $\{\mathrm{C}, \mathrm{D}\}$. The state $(\mathrm{C}, \mathrm{C}, \mathrm{C})$ is in the neighbouring set of $(\mathrm{C}, \mathrm{C}, \mathrm{D})$ whereas the state $(\mathrm{C}, \mathrm{D}, \mathrm{D})$ is not in the neighbourhood set of $(\mathrm{C}, \mathrm{C}, \mathrm{C})$. Two states that belong in each other's neighbourhood set only differ in exactly a single player's (that we call as the index of difference) action.

Definition 2 (Index of difference between neighbouring states)**.** *If two states,* $\mathbf{a}$ *and* $\mathbf{b}$*, satisfy* $\mathbf{a} \in \mathrm{Neb}(\mathbf{b})$*, the index of difference between them,* $\mathrm{I}(\mathbf{a}, \mathbf{b})$*, is the unique integer that satisfies:*

$$\mathbf{a}_{\mathrm{I}(\mathbf{a},\mathbf{b})} \neq \mathbf{b}_{\mathrm{I}(\mathbf{a},\mathbf{b})} \tag{4}$$

In the previous example, the index of difference between the neighbouring states $(\mathrm{C}, \mathrm{C}, \mathrm{C})$ and $(\mathrm{C}, \mathrm{C}, \mathrm{D})$ is 3. The third player's action is the difference between the two neighbouring states. Using the above definitions, one can formally define the transition matrix of the introspection dynamics with:

$$\mathbf{T}_{\mathbf{a},\mathbf{b}} = \begin{cases} \frac{1}{N(m_j - 1)} \cdot p_{\mathbf{a}_j \to \mathbf{b}_j}(\mathbf{a}_{-j}) & \text{if } \mathbf{b} \in \mathrm{Neb}(\mathbf{a}) \quad \text{and,} \quad j = \mathrm{I}(\mathbf{a}, \mathbf{b}) \\ 0 & \text{if } \mathbf{b} \notin \mathrm{Neb}(\mathbf{a}) \\ 1 - \sum_{\mathbf{c} \neq \mathbf{b}} \mathbf{T}_{\mathbf{a},\mathbf{c}} & \text{if } \mathbf{a} = \mathbf{b} \end{cases} \tag{5}$$

The transition matrix is a row stochastic matrix (the sums of the rows are 1). This implies that the stationary distribution of $\mathbf{T}$: a left eigenvector of $\mathbf{T}$ corresponding to eigenvalue 1, always exists. The following proposition introduces a sufficient condition for this stationary distribution to be unique.

Proposition 1. *When* β *is finite, the transition matrix of the introspection dynamics has a unique stationary distribution.*

Proof: A finite value of β results in non-zero probability of transition between neighbouring states. Since no state is isolated (i.e., every state belongs in the neighbourhood set of another state), every state is reachable in a finite number of steps from a starting point with non-zero probability. The transition

matrix $\mathbf{T}$ is therefore primitive for a finite β. By the Perron-Frobenius theorem, a primite matrix, $\mathbf{T}$ will have a unique and strictly positive stationary distribution $\mathbf{u} := (\mathbf{u_a})_{\mathbf{a}\in\mathbf{A}}$ which satisfies the conditions:

$$\mathbf{uT} = \mathbf{u} \tag{6}$$

$$\mathbf{u1} = 1 \tag{7}$$

where $\mathbf{1}$ is the column vector with size same as $\mathbf{u}$ and has all elements as 1. The above equations only present an explicit representation of the stationary distribution $\mathbf{u}$. The stationary distribution can be explictly calculated by the following expression (which is derived using Eq. (6) and (7) as:

$$\mathbf{u} = \mathbf{1}^{\intercal}(\mathbb{1} + \mathbf{U} - \mathbf{T})^{-1} \tag{8}$$

where $\mathbf{U}$ is a square matrix of size same as $\mathbf{T}$ with all elements 1 and $\mathbb{1}$ is the identity matrix. For all the analytical results in this paper, we consider β to be finite so that stationary distribution of the processes are unique. The stationary distribution element $\mathbf{u_a}$ is the probability that state $\mathbf{a}$ will be played by the players in the long run. Using the stationary distribution, one can calculate the marginal probabilities corresponding to each player's actions. For example, the probability that player i plays action $a_{i,k}$ in the long run, $\xi_{i:a_{i,k}}$, can be computed as,

$$\xi_{i:a_{i,k}} := \sum_{\mathbf{q}\in\mathbf{A}_{-i}} \mathbf{u}_{(a_{i,k},\mathbf{q})} \tag{9}$$

Since the stationary distribution is a probability distribution, marginal distributions also have the same property. That is, for a player i,

$$\sum_{k=1}^{m_i} \xi_{i:a_{i,k}} = 1 \tag{10}$$

Additive games and their properties under introspection dynamics

In this section we discuss the stationary properties of introspection dynamics on a special class of games - the additive games[2,3]. In an additive game, the payoff difference that a player earns by making a switch in their actions is independent of what their co-players play while they make this switch. In other words, if none of the co-players change their strategy, the payoff difference earned by making a particular switch is *only* determined by the switch and not on the co-players' play. This property is sometimes called the *equal gains from switching*[2] property of a game. That is, for any player i, any pair of actions $x, y \in \mathbf{A}_i$, and any $\mathbf{q} \in A_{-i}$,

$$\pi_i(x, \mathbf{q}) - \pi_i(y, \mathbf{q}) =: f_i(x, y) \tag{11}$$

For games with this property, the stationary distribution takes a simple form as shown in the proposition below.

Proposition 2. *When β is finite, the unique stationary distribution, $\mathbf{u} = (\mathbf{u_a})_{\mathbf{a}\in\mathbf{A}}$, of the introspection dynamics for the N-player additive game is given by:*

$$\mathbf{u_a} = \prod_{j=1}^{N} \frac{1}{\sum\limits_{a'\in\mathbf{A}_j} e^{\beta f_j(a',\mathbf{a}_j)}} \tag{12}$$

where, $f_j(a',\mathbf{a}_j) = \pi_j(a',\mathbf{q}) - \pi_j(\mathbf{a}_j,\mathbf{q})$ is the co-player independent payoff difference that j earns in an additive game when they unilaterally switch their play to a' from $\mathbf{a}_j$.

The above proposition gives a closed form expression of the stationary distribution of the introspection dynamics for additive games. For any finite value of the selection parameter, β, the stationary distribution can be computed using the expression in Eq. (12). For proof, see Appendix . Additive games are particularly interesting under the introspection dynamics because of the following relationship between the stationary distribution and marginal distributions,

Proposition 3. *Let $\mathbf{u} = (\mathbf{u_a})_{\mathbf{a}\in\mathbf{A}}$ be the unique stationary distribution of the introspection dynamics with finite β for the N-player additive game. Then, $\mathbf{u_a}$ is the product of the marginal probabilities that each player plays their respective actions in $\mathbf{a}$. That is,*

$$\mathbf{u_a} = \prod_{j=1}^{N} \xi_{j:\mathbf{a}_j} \tag{13}$$

where the marginal probability, $\xi_{j:\mathbf{a}_j}$, is the cumulative probability that player j plays $\mathbf{a}_j$ at the stationary distribution. The marginal probability is given by:

$$\xi_{j:\mathbf{a}_j} = \frac{1}{\sum\limits_{a'\in\mathbf{A}_j} e^{\beta f_j(a',\mathbf{a}_j)}} \tag{14}$$

where, $f_j(a',\mathbf{a}_j) = \pi_j(a',\mathbf{q}) - \pi_j(\mathbf{a}_j,\mathbf{q})$ is the co-player independent payoff difference that j earns in an additive game when they unilaterally switch their play to a' from $\mathbf{a}_j$.

The above proposition states that for additive games under introspection dynamics, the stationary distribution can be factorized into its corresponding marginals. In the long run, the probability that players simultaneously play $(\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N)$ is the product of the cumulative probabilities that player 1 plays $\mathbf{a}_1$, player 2 plays $\mathbf{a}_2$ and so on. In the following subsection we study a classical additve game - the linear public goods game under the introspection dynamics and discuss the results.

Example of an additive game: linear public goods game with 2 actions

Here, we analyze the linear public goods game with N-players. Each player has two possible actions, to contribute (action, C), or to not contribute (action, D). The players differ in their cost of cooperation and the benefit they provide to the public goods. The cost of cooperation for player i and the benefit they provide to the public goods are denoted with c_i and b_i respectively. The payoff of player i when the state of the game is $\mathbf{a}$ is given by:

$$\pi_i(\mathbf{a}) = \sum_{j=1}^{N} \frac{\mathbf{a}_j b_j}{N} - \mathbf{a}_i c_i \tag{15}$$

The payoff difference that a player earns by unilaterally switching from C to D (or *vice-versa*) in the linear public goods game is independent of what the other co-players play in the game. That is, for every player i,

$$\pi_i(\mathrm{D}, \mathbf{q}) - \pi_i(\mathrm{C}, \mathbf{q}) = c_i - \frac{b_i}{N} =: f_i(\mathrm{D}, \mathrm{C}) \tag{16}$$

is independent of any $\mathbf{q} \in \mathbf{A}_{-i}$ that the co-players of i play. The linear public goods game is therefore an example of an additive game. This property of the game results in dominated strategies for player i depending on the relationship between the benefits b_i, number of players N, and cost of cooperation c_i. When $c_i > b_i/N$, the action of not contributing (D) dominates the action of contributing (C) and *vice-versa*. Using proposition 2, we derive the following closed form expression for the stationary distribution of a $N-$player linear public goods game with two strategies.

Proposition 4. *When β is finite, the unique stationary distribution of the introspection dynamics for a $N-$player linear public goods game is given by:*

$$\mathbf{u}_{\mathbf{a}} = \prod_{j=1}^{N} \frac{1}{1 + e^{sign(\mathbf{a}_j)\beta f_j(\mathrm{D},\mathrm{C})}} \tag{17}$$

where,

$$sign(a) = \begin{cases} 1 & \text{if} \quad a = \mathrm{C} \\ -1 & \text{if} \quad a = \mathrm{D} \end{cases} \tag{18}$$

Appendix: Proofs

Proof. **Proof of Proposition** 2

Since β is finite,the stationary distribution $\mathbf{u} = (\mathbf{u_a})_{\mathbf{a}\in\mathbf{A}}$ of the process is unique by Proposition 1. The stationary distribution also satisfies the equalities in Eq. (6) and (7). Before continuing through the remainder of the proof, we introduce some short-cut notation that we will be using:

$$\mathrm{I}_{\mathbf{b}} := \mathrm{I}(\mathbf{b}, \mathbf{a}), \quad \textit{iff} \quad \mathbf{b} \in \mathrm{Neb}(\mathbf{a}) \tag{19}$$

$$\tau_{j:\mathbf{a}_j} := \frac{1}{\sum\limits_{a' \in \mathbf{A}_j} e^{\beta f_j(a', \mathbf{a}_j)}} \tag{20}$$

In order to show that the candidate stationary distribution, as proposed in Eq. (12) is the stationary distribution of the process, we need to show that the following are true:

$$\mathbf{T}_{\mathbf{a},\mathbf{a}}\mathbf{u}_{\mathbf{a}} + \sum_{\mathbf{b}\neq\mathbf{a}} \mathbf{T}_{\mathbf{b},\mathbf{a}}\mathbf{u}_{\mathbf{b}} = \mathbf{u}_{\mathbf{a}} \quad \forall \mathbf{a} \in \mathbf{A} \tag{21}$$

$$\sum_{\mathbf{a}\in\mathbf{A}} \mathbf{u}_{\mathbf{a}} = 1 \tag{22}$$

Using our short-cut notation τ and the expression for our candidate stationary distribution in Eq. (12), we can express the stationary distribution as:

$$\mathbf{u}_{\mathbf{a}} = \prod_{j=1}^{N} \tau_{j:\mathbf{a}_j} \tag{23}$$

Using this expression, the left hand side of Eq. (21) can be simplified further with the steps:

$$\mathbf{T}_{\mathbf{a},\mathbf{a}}\mathbf{u}_{\mathbf{a}} + \sum_{\mathbf{b}\neq\mathbf{a}} \mathbf{T}_{\mathbf{b},\mathbf{a}}\mathbf{u}_{\mathbf{b}} \tag{24}$$

$$= \left(1 - \frac{1}{N} \sum_{\mathbf{b}\in\mathrm{Neb}(\mathbf{a})} \frac{1}{m_{\mathrm{I}_{\mathbf{b}}} - 1} \cdot p_{\mathbf{a}_{\mathrm{I}_{\mathbf{b}}} \to \mathbf{b}_{\mathrm{I}_{\mathbf{b}}}} \cdot \mathbf{u}_{\mathbf{a}}\right) + \sum_{\mathbf{b}\in\mathrm{Neb}(\mathbf{a})} \frac{1}{m_{\mathrm{I}_{\mathbf{b}}} - 1} \cdot p_{\mathbf{b}_{\mathrm{I}_{\mathbf{b}}} \to \mathbf{a}_{\mathrm{I}_{\mathbf{b}}}} \cdot \mathbf{u}_{\mathbf{b}} \tag{25}$$

$$= \mathbf{u}_{\mathbf{a}} + \frac{1}{N} \sum_{\mathbf{b}\in\mathrm{Neb}(\mathbf{a})} \left(\prod_{k\neq I_{\mathbf{b}}} \tau_{k:\mathbf{a}_k}\right) \left(p_{\mathbf{b}_{I_{\mathbf{b}}} \to \mathbf{a}_{\mathrm{I}_{\mathbf{b}}}} \cdot \tau_{\mathrm{I}_{\mathbf{b}}:\mathbf{a}_{\mathrm{I}_{\mathbf{b}}}} - p_{\mathbf{a}_{\mathrm{I}_{\mathbf{b}}} \to \mathbf{b}_{\mathrm{I}_{\mathbf{b}}}} \cdot \tau_{\mathrm{I}_{\mathbf{b}}:\mathbf{b}_{\mathrm{I}_{\mathbf{b}}}}\right) \cdot \left(\frac{1}{m_{\mathrm{I}_{\mathbf{b}}} - 1}\right) \tag{26}$$

For an additive game, the expressions for $p_{\mathbf{b}_{I_\mathbf{b}} \to \mathbf{a}_{\mathrm{I}_\mathbf{b}}}$ and $p_{\mathbf{a}_{\mathrm{I}_\mathbf{b}} \to \mathbf{b}_{\mathrm{I}_\mathbf{b}}}$ can be simply written as:

$$p_{\mathbf{b}_{I_\mathbf{b}} \to \mathbf{a}_{\mathrm{I}_\mathbf{b}}} = \frac{1}{1 + e^{\beta f_{\mathrm{I}_\mathbf{b}}(\mathbf{b}_{\mathrm{I}_\mathbf{b}}, \mathbf{a}_{\mathrm{I}_\mathbf{b}})}} \tag{27}$$

$$p_{\mathbf{a}_{\mathrm{I}_\mathbf{b}} \to \mathbf{b}_{\mathrm{I}_\mathbf{b}}} = \frac{1}{1 + e^{\beta f_{\mathrm{I}_\mathbf{b}}(\mathbf{a}_{\mathrm{I}_\mathbf{b}}, \mathbf{b}_{\mathrm{I}_\mathbf{b}})}} \tag{28}$$

Using the above expressions and the expression for τ in Eq. (20), it can be shown that:

$$\left(p_{\mathbf{b}_{I_\mathbf{b}} \to \mathbf{a}_{\mathrm{I}_\mathbf{b}}} \cdot \tau_{\mathrm{I}_\mathbf{b}:\mathbf{a}_{\mathrm{I}_\mathbf{b}}} - p_{\mathbf{a}_{\mathrm{I}_\mathbf{b}} \to \mathbf{b}_{\mathrm{I}_\mathbf{b}}} \cdot \tau_{\mathrm{I}_\mathbf{b}:\mathbf{b}_{\mathrm{I}_\mathbf{b}}} \right) = 0 \tag{29}$$

After plugging the equality in Eq. (29) into Eq. (26), we see that the left hand side of Eq. (21) simplifies to $\mathbf{u_a}$. Now, to complete the proof we must check if Eq. (22) holds for our candidate distribution. Summing up the elements of the stationary distribution $\mathbf{u_a}$ for all states $\mathbf{a} \in \mathbf{A}$:

$$\sum_{\mathbf{a} \in \mathbf{A}} \mathbf{u_a} = \sum_{\mathbf{a} \in \mathbf{A}} \prod_{k=1}^{N} \tau_{k:\mathbf{a}_k} = \sum_{\mathbf{a} \in \mathbf{A}} \frac{\prod_{k=1}^{N} e^{\beta \pi_k(\mathbf{a}_k, \mathbf{q}_{-k})}}{\prod_{k=1}^{N} \sum_{a' \in \mathbf{A}_k} e^{\beta \pi_k(a', \mathbf{q}_{-k})}} \tag{30}$$

where $\mathbf{q}_{-1}, \mathbf{q}_{-2}, ..., \mathbf{q}_{-N}$ are fixed tuples from $\mathbf{A}_{-1}, \mathbf{A}_{-2}, ..., \mathbf{A}_{-N}$ respectively. The denominator in the above expression can be taken out completely from the first sum. That is,

$$\sum_{\mathbf{a} \in \mathbf{A}} \mathbf{u_a} = \sum_{\mathbf{a} \in \mathbf{A}} \frac{\prod_{k=1}^{N} e^{\beta \pi_k(\mathbf{a}_k, \mathbf{q}_{-k})}}{\prod_{k=1}^{N} \sum_{a' \in \mathbf{A}_k} e^{\beta \pi_k(a', \mathbf{q}_{-k})}} \tag{31}$$

$$= \left(\prod_{k=1}^{N} \left(e^{\beta \pi_k(a_{k,1}, \mathbf{q}_{-k})} + ... + e^{\beta \pi_k(a_{k,m_k}, \mathbf{q}_{-k})} \right) \right)^{-1} \cdot \left(\sum_{\mathbf{a} \in \mathbf{A}} \prod_{k=1}^{N} e^{\beta \pi_k(\mathbf{a}_k, \mathbf{q}_{-k})} \right) \tag{32}$$

$$\tag{33}$$

Multiplying out the sums in the denominator of the above expression, we get that:

$$\sum_{\mathbf{a}\in\mathbf{A}} \mathbf{u}_{\mathbf{a}} = \left(\prod_{k=1}^{N}\left(e^{\beta\pi_k(a_{k,1},\mathbf{q}_{-k})} + ... + e^{\beta\pi_k(a_{k,m_k},\mathbf{q}_{-k})}\right)\right)^{-1} \cdot \left(\sum_{\mathbf{a}\in\mathbf{A}}\prod_{k=1}^{N} e^{\beta\pi_k(\mathbf{a}_k,\mathbf{q}_{-k})}\right) \tag{34}$$

$$= \left(\sum_{\mathbf{a}\in\mathbf{A}}\prod_{k=1}^{N} e^{\beta\pi_k(\mathbf{a}_k,\mathbf{q}_{-k})}\right)^{-1}\left(\sum_{\mathbf{a}\in\mathbf{A}}\prod_{k=1}^{N} e^{\beta\pi_k(\mathbf{a}_k,\mathbf{q}_{-k})}\right) = 1 \tag{35}$$

□

Proof. **Proof of Proposition** 3

If $\mathbf{u}$ is the unique stationary distribution of the $N-$player additive game with under the finite selection introspection dynamics, it is given by the expression in Eq. 12. We calculate the marginal distribution of any arbitrary state $\mathbf{a}$, $\xi_{\mathbf{a}} = (\xi^j_{\mathbf{a}^j})_{j=1,2,...,N}$ by using the definition of marginal distribution in Eq. 9. It follows that:

$$\xi^j_{\mathbf{a}^j} = \sum_{\mathbf{b}\in\mathbf{A}^{-j}} \mathbf{u}_{(\mathbf{a}^j,\mathbf{b})} \tag{36}$$

$$= \left(\prod_{k=1}^{N}\sum_{\mathbf{a}'\in\mathbf{A}^k} e^{\beta\pi^k_{\mathbf{a}'}}\right)^{-1} \cdot e^{\beta\pi^j_{\mathbf{a}^j}} \cdot \left(\sum_{\mathbf{b}\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{b}^k}}\right) \tag{37}$$

$$= \left(\sum_{\mathbf{a}'\in\mathbf{A}^j} e^{\beta\pi^j_{\mathbf{a}'}}\right)^{-1} \cdot e^{\beta\pi^j_{\mathbf{a}^j}} \cdot \left(\prod_{k\neq j}\sum_{\mathbf{a}'\in\mathbf{A}^k} e^{\beta\pi^k_{\mathbf{a}'}}\right)^{-1} \cdot \left(\sum_{\mathbf{b}\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{b}^k}}\right) \tag{38}$$

$$= \left(\sum_{\mathbf{a}'\in\mathbf{A}^j} e^{\beta\pi^j_{\mathbf{a}'}}\right)^{-1} \cdot e^{\beta\pi^j_{\mathbf{a}^j}} \cdot \left(\sum_{\mathbf{a}'\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{a}'}}\right)^{-1} \cdot \left(\sum_{\mathbf{b}\in\mathbf{A}^{-j}}\prod_{k\neq j} e^{\beta\pi^k_{\mathbf{b}^k}}\right) \tag{39}$$

$$= \left(\sum_{\mathbf{a}'\in\mathbf{A}^j} e^{\beta\left(\pi^j_{\mathbf{a}'} - \pi^j_{\mathbf{a}^j}\right)}\right)^{-1} \tag{40}$$

Therefore, the marginal distribution follows Eq. 14. Now since we also additionally know that the stationary distribution follows the form Eq. 12, we can conclude that for additive games, under introspection dynamics with finite selection, Eq. 13 holds. □

Proof. **Proof of Proposition** 4

Since we have demonstrated that the linear public goods game is an additive game, the proof of this theorem can be performed by directly using Theorem 2. Here, we provide an independent proof. The idea behind this proof is identical to the proof of Theorem 2.

The stationary transition matrix $\mathbf{T}$ for the linear public goods game is primitive when β is finite (i.e., there is a positive power k such that $\mathbf{T}^k$ is a strictly positive matrix). Therefore, the stationary distribution of $\mathbf{T}$ will always be unique. We define the following short cut notations for the ease of the proof:

$$\bar{\mathbf{a}}^j := \{\mathrm{D}, \mathrm{C}\} \setminus \mathbf{a}^j \tag{41}$$

$$p^j := \frac{1}{1+ e^{\beta f(c^j, r^j)}} \tag{42}$$

In addition we introduce a mapping function $\alpha(.)$ which maps the action C to 1 and the action D to 0. That is $\alpha(\mathrm{C}) := 1$ and $\alpha(\mathrm{D}) := 0$. Using these notations and Eq. 1 and 16 we can write the probability that a player j updates from $\mathbf{a}^j$ to $\bar{\mathbf{a}}^j$ while their co-players play $\mathbf{a}^{-j}$ as:

$$p^j_{\bar{\mathbf{a}}^j \rightarrow \mathbf{a}^j}(\mathbf{a}^{-j}) = p^j sign(\mathbf{a}^j) + \alpha(\bar{\mathbf{a}}^j) \tag{43}$$

The candidate stationary distribution $\mathbf{u}$ given in Eq 17 can be written down using our shortcut notation as:

$$\mathbf{u}_{\mathbf{a}} = \prod_{k=1}^{N} p^k sign(\mathbf{a}^k) + \alpha(\bar{\mathbf{a}}^k) \quad , \forall \mathbf{a} \in \{0,1\}^N \tag{44}$$

This stationary distribution must satisfy the following properties, which are also given in Eq 6 and 7:

$$\mathbf{u}_{\mathbf{a}} = \mathbf{T}_{\mathbf{a},\mathbf{a}}\mathbf{u}_{\mathbf{a}} + \sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}}\mathbf{u}_{\mathbf{a}_q} \tag{45}$$

$$\sum_{\forall \mathbf{a}_q} \mathbf{u}_{\mathbf{a}_q} = 1 \tag{46}$$

Where, the terms in the right hand side of Eq. 45 can be simplified using Eq 1 and 5 as follows:

$$\mathbf{T}_{\mathbf{a},\mathbf{a}} = 1 - \sum_{k=1}^{N} \mathbf{T}_{(\mathbf{a}^k,\mathbf{a}^{-k}),(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} = 1 - \frac{1}{N}\sum_{k=1}^{N} p^k sign(\bar{\mathbf{a}}^k) + \alpha(\mathbf{a}^k) \tag{47}$$

and additionally, also using Eq. 44 the second term can be simplified too:

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \mathbf{u}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} \tag{48}$$

$$= \frac{1}{N} \sum_{k=1}^{N} \left(p^k sign(\mathbf{a}^k) + \alpha(\bar{\mathbf{a}}^k) \right) \mathbf{u}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k})} \tag{49}$$

$$= \frac{\mathbf{u_a}}{N} \sum_{k=1}^{N} p^k sign(\bar{\mathbf{a}}^k) + \alpha(\mathbf{a}^k) \tag{50}$$

Now, using Eq. 47, 50 one can show that the right hand side of Eq. 45 is the element of the stationary distribution, corresponding to the state $\mathbf{a}$, $\mathbf{u}_a$. Now, to complete the proof, we must show that Eq. 46 is also true for our candidate stationary distribution. This can be done by decomposing the sum of the elements of the stationary distribution as follows:

$$\sum_{\forall \mathbf{a}_q} u_{\mathbf{a}_q} = \sum_{\forall \mathbf{a}_q} \prod_{k=1}^{N} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) \tag{51}$$

$$= \sum_{\forall \mathbf{a}_q} (1 - p^N) \prod_{k=1}^{N-1} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) + p^N \prod_{k=1}^{N-1} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) \tag{52}$$

$$= \sum_{\forall \mathbf{a}_q} \prod_{k=1}^{N-1} p^k sign(\mathbf{a}_q^k) + \alpha(\bar{\mathbf{a}}_q^k) \tag{53}$$

When the above decomposition is perfomed $N - 1$ more times, the sum of the right hand side becomes 1. This prooves that the candidate stationary distribution is also a probability distribution. □

Proof. **Proof of Proposition ??**

By construction, the candidate stationary distribution given by Eq. **??** and Eq. **??** is a probability distribution since it satisfies the condition in Eq. 7 and for any state $\mathbf{a}'$, $\mathbf{u}_{\mathbf{a}'}$ is between 0 and 1. Moreover, since β is finite, the transition matrix of the process $\mathbf{T}$ is primitive and therefore, it will have a unique stationary distribution. To show that the candidate stationary distribution is the unique stationary distribution, we need to check if for this process, $\mathbf{uT} = \mathbf{u}$. That is, the condition in Eq. 45 must hold for

all states $\mathbf{a}$. We re-introduce some notations that we will use in this proof:

$$\bar{\mathbf{a}}^j := \{\mathrm{D}, \mathrm{C}\} \setminus \mathbf{a}^j \tag{54}$$

$$\alpha(a) := \begin{cases} 1 & \text{if} \quad a = \mathrm{C} \\ 0 & \text{if} \quad a = \mathrm{D} \end{cases} \tag{55}$$

$$\mathcal{C}(\mathbf{a}) = \sum_{j=1}^{N} \alpha(\mathbf{a}^j) \tag{56}$$

For this process, the first term in the right hand side of Eq. 45 can be simplified as:

$$\mathbf{u_a}\mathbf{T_{a,a}} = \mathbf{u_a} - \mathbf{u_a} \sum_{k=1}^{N} \mathbf{T}_{(\mathbf{a}^k, \mathbf{a}^{-k}), (\bar{\mathbf{a}}^k, \mathbf{a}^{-k})} \tag{57}$$

$$= \mathbf{u_a} - \frac{\mathbf{u_a}}{N} \sum_{k=1}^{N} \frac{1}{1 + e^{sign(\bar{\mathbf{a}}^k)\beta f(N_k)}} \tag{58}$$

Where, the function $sign(.)$ is defined as in Eq. 18 and $f(j)$ is the difference in payoffs between playing D and C when there are j co-players playing C. The term N_k is the number of co-players of k that play C in state $\mathbf{a}$. That is,

$$N_k := \sum_{j \neq k} \alpha(\mathbf{a}^j) \tag{59}$$

The second term in the right hand side of Eq. 45 can be simplified as,

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q, \mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k, \mathbf{a}^{-k}), (\mathbf{a}^k, \mathbf{a}^{-k})} \mathbf{u}_{(\bar{\mathbf{a}}^k, \mathbf{a}^{-k})} \tag{60}$$

$$= \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k, \mathbf{a}^{-k}), (\mathbf{a}^k, \mathbf{a}^{-k})} \prod_{j=1}^{\mathcal{C}((\bar{\mathbf{a}}^k, \mathbf{a}^{-k}))} e^{-\beta f(j-1)} \tag{61}$$

$$= \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k, \mathbf{a}^{-k}), (\mathbf{a}^k, \mathbf{a}^{-k})} \left(\prod_{j=1}^{N_k} e^{-\beta f(j-1)} \right) \cdot e^{-\beta \alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k) + N_k} \tag{62}$$

Between the steps, Eq. 61 and 62, we took out one term from the product that is present in our candidate distribution. This term accounts for the k^{th} players action in the neighbouring state $(\bar{\mathbf{a}}^k, \mathbf{a}^{-k})$ of $\mathbf{a}$. For simplicity, we replace $\mathbf{T}_{(\bar{\mathbf{a}}^k, \mathbf{a}^{-k}), (\mathbf{a}^k, \mathbf{a}^{-k})}$ with just $\mathbf{T}$ in the next steps. We continue the simplification of Eq. 62 in the later steps by introducing terms that cancel each other.

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T} \cdot \left(\prod_{j=1}^{N_k} e^{-\beta f(j-1)} \right) \cdot \frac{e^{-\beta\alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k)+N_k)}}{e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)}} \cdot e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)} \quad (63)$$

The newly introduced term in Eq. 63 can be taken inside the product. Note that this term is 0 if the k^{th} player plays D in the state $\mathbf{a}$. When this term is taken inside the product bracket, products of exponent $e^{-\beta f(j-1)}$ can be performed for j ranging from 1 to the number of cooperators in state $\mathbf{a}$, $\mathcal{C}(\mathbf{a})$. This product is then the stationary distribution probability $\mathbf{u}_{\mathbf{a}}$. That is,

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{1}{N\Gamma} \sum_{k=1}^{N} \mathbf{T} \cdot \left(\prod_{j=1}^{N_k} e^{-\beta f(j-1)} \cdot e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)} \right) \cdot \frac{e^{-\beta\alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k)+N_k)}}{e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)}} \quad (64)$$

$$= \frac{1}{N} \sum_{k=1}^{N} \mathbf{T} \cdot \left(\frac{1}{\Gamma} \prod_{j=1}^{\mathcal{C}(\mathbf{a})} e^{-\beta f(j-1)} \right) \cdot \frac{e^{-\beta\alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k)+N_k)}}{e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)}} \quad (65)$$

$$= \frac{1}{N} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \cdot \mathbf{u}_{\mathbf{a}} \cdot \frac{e^{-\beta\alpha(\bar{\mathbf{a}}^k) f(-\alpha(\mathbf{a}^k)+N_k)}}{e^{-\beta\alpha(\mathbf{a}^k) f(-\alpha(\bar{\mathbf{a}}^k)+N_k)}} \quad (66)$$

The fraction inside the sum in Eq. 66 can be simplified as follows leading to further simplification of Eq. 66:

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} \cdot \mathbf{u}_{\mathbf{a}} \cdot e^{sign(\mathbf{a}^k)\beta f(N_k)} \quad (67)$$

In Eq. 67 we can replace the element of the transition matrix $\mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})}$ by using the following:

$$\mathbf{T}_{(\bar{\mathbf{a}}^k,\mathbf{a}^{-k}),(\mathbf{a}^k,\mathbf{a}^{-k})} = \frac{1}{1 + e^{sign(\mathbf{a}^k)\beta f(N_k)}} \quad (68)$$

Using the expression for the transition matrix element from Eq. 68 into Eq. 67 and by using Eq. 58, we

can simplify further:

$$\sum_{\mathbf{a}_q \neq \mathbf{a}} \mathbf{T}_{\mathbf{a}_q,\mathbf{a}} \mathbf{u}_{\mathbf{a}_q} = \frac{\mathbf{u_a}}{N} \sum_{k=1}^{N} \frac{1}{1 + e^{sign(\mathbf{a}^k)\beta f(N_k)}} \cdot e^{sign(\mathbf{a}^k)\beta f(N_k)} \tag{69}$$

$$= \frac{\mathbf{u_a}}{N} \sum_{k=1}^{N} \frac{1}{1 + e^{sign(\bar{\mathbf{a}}^k)\beta f(N_k)}} \tag{70}$$

$$= \mathbf{u_a} - \mathbf{u_a} \mathbf{T}_{\mathbf{a},\mathbf{a}} \tag{71}$$

The final step in the previous simplification shows that Eq. 45 holds for any $\mathbf{a} \in \{\mathrm{C}, \mathrm{D}\}^N$. Therefore, the candidate distribution we propose in Eq. **??** is the unique stationary distribution of the symmetric N-player game with two strategies. □

Proof. **Proof of Corollary ??**

To show this result we count how many states are identical to a state $\mathbf{a} \in \{\mathrm{C}, \mathrm{D}\}^N$ in a symmetric game. When players are symmetric in a two-strategy game, states can be enumerated by counting the number of C players in that state. This can also be confirmed by the expression of the stationary distribution in Eq. **??**. Two distinct states $\mathbf{a}, \mathbf{a}'$ having the same number of cooperators (i.e., $\mathcal{C}(\mathbf{a}') = \mathcal{C}(\mathbf{a})$), have the same stationary distribution probability (i.e., $\mathbf{u}_{\mathbf{a}'} = \mathbf{u}_{\mathbf{a}}$).

In a game with N players, there can be k players playing C in exactly $\binom{N}{k}$ ways. As argued before, all of these states are identical and are also equiprobable in the stationary distribution. Therefore, the stationary distribution probability of having k, C players, $\mathbf{u}_k$, is,

$$\mathbf{u}_k = \sum_{\forall \mathbf{a}, \mathcal{C}(\mathbf{a})=k} \mathbf{u}_{\mathbf{a}} = \frac{1}{\Gamma} \binom{N}{k} \prod_{j=1}^{k} e^{-\beta f(j-1)} \tag{72}$$

Where the normalization factor Γ can also be simplified as:

$$\Gamma = \sum_{k=0}^{N} \binom{N}{k} \prod_{j=1}^{k} e^{-\beta f(j-1)} \tag{73}$$

□

Supplementary References

References

[1] Marta Couto, Stefano Giaimo, and Christian Hilbe. Introspection dynamics: A simple model of counterfactual learning in asymmetric games. *New Journal of Physics*, 2022.

[2] Jorge Pena, Laurent Lehmann, and Georg Nöldeke. Gains from switching and evolutionary stability in multi-player matrix games. *Journal of Theoretical Biology*, 346:23–33, 2014.

[3] Alex McAvoy and Christoph Hauert. Asymmetric evolutionary games. *PLoS computational biology*, 11(8):e1004349, 2015.