Evolution of cooperation among individuals with limited payoff memory

Christian Hilbe, Nikoleta E. Glynatsi, Alex McAvoy

Abstract

1 Introduction

One of the most important applications of evolutionary game theory is the evolution of cooperation. Why is it that some individuals choose to help others, increasing their payoff, at the expense of decreasing one's own? In evolutionary game theory, individuals are not required to be rational, instead they adapt strategies based on mutation and exploration. Strategies are more likely to spread if they have a high fitness either because the individuals who adopt them have more offspring, or because they are imitated more often. The fitness of a strategy is not constant but depends on the composition of the population. Individuals interact based on their strategies with other members of the population and the payoffs they yield are translated into fitness.

It is commonly assumed that fitness is equivalent to the mean distribution of types in the population. These payoffs assume that individuals can interact with the entire population several times and remember each and every outcome. Thus, they imply that individuals have a perfect memory. However, when they make decisions in each turn they are assumed to have very limited memory. To be precise, most of the works in the literature focus on naive subjects who can only choose from a restricted set of strategies [1], or who do not remember anything beyond the outcome of the very last round [2]. Note that there are a few notable exceptions [3, 4].

The perfect memory assumption is not only unrealistic but it also creates this curious inconsistency. This has led us to question how robust is our understanding of cooperation. In this work, we explore whether direct reciprocity can evolve if individuals only remember a minimum of social information. Though we are not the first to question the assumptions of estimating fitness [5], we are the first to explore the effect of payoff memory.

Initially, we consider two extreme scenarios. The first is the classical scenario of the expected payoffs and the alternative scenario where individuals update their strategies only based on the very last payoff they obtained. In the later sections, we allow individuals more memory. More specifically, they can remember up to two turns and up to two interactions. We present results on several social dilemmas which include the prisoner's dilemma and the donation game, the snowdrift game, the stag hunt game and the harmony game.

The remainder of the paper is organized as follows. Section 2 describes the model. Section 3 presents the results of the simulations. Finally, section 4 outlines the main conclusions.

2 Model Setup

We consider a population of N players (N is even) where mutations are sufficiently rare. Thus, at any point in time there are at most two different strategies present in the population; a *resident* strategy and a *mutant* strategy. We assume a pairwise process where strategies spread because they are imitated more often.

Each step of the evolutionary process consists of two stages, a game stage and an updating stage. In the game stage each individual is randomly matched with some other individual in the population to interact for a number of turns, where subsequent turns occur with a fixed probability δ. At each turn they can choose to either cooperate (C) or to defect (D), and thus, at each turn the possible outcomes are CC, CD, DC and DD. The payoffs depend on the outcome. If both cooperate they receive the reward payoff R, whereas if both defect they receive the punishment payoff P. If one cooperates but the other defects, the defector receives the temptation to defect, T, whereas the cooperator receives the sucker's payoff, S. We denote the payoffs of an individual as $\mathcal{U} = (R, S, T, P)$.

We assume herein that individuals use *reactive strategies* to make decisions in each turn. Reactive strategies are a set of memory-one strategies that only take into account the previous action of the opponent. They can be written explicitly as a vector $\in \mathbb{R}_3$, more specifically, a reactive strategy s is given by $s = (y, p, q)$ where y is the probability that the strategy opens with a cooperation and p, q are the probabilities that the strategy cooperates given that the opponent cooperated and defected equivalently.

In the updating stage, two players are randomly drawn from the population, a 'learner' and a 'role model'. Given that the learner's payoff $u_L \in \mathcal{U}$ and that the role model's payoff $u_{RL} \in \mathcal{U}$, we assume the learner adopts the role model's strategy based on the Fermi distribution function,

$$\rho(u_L, u_{RM}) = \frac{1}{1+\exp^{-\beta(u_{RM}-u_L)}}. \quad (1)$$

where $\beta \geq 0$ is the relative influence of the payoffs on adopting the strategy of the other. We refer to β as the intensity of selection.

This basic evolutionary step is repeated until either the mutant strategy goes extinct, or until it fixes in the population. If the mutant fixes in the population then the mutant strategy becomes the new resident strategy. After either outcome we introduce a new mutant strategy uniformly chosen from all reactive strategies at random, and we set the number of mutants to 1. This process of mutation and fixation/extinction is then iterated many times.

The perfect memory assumption occurs at the updating stage. The learner and the role model are assumed to interact with a representative sample of the population, and they remember all interactions they participate in. Thus, their updating payoffs are based on the mean payoff they achieved over all the interactions. These payoffs are referred to as the expected payoffs. We will compare the expected payoffs to payoffs that are calculated when the role model and learner do not remember all of their interactions. In order to account for the effect of these different methods, we explore the cooperation rate within the resident population over multiple generations. More details on our methodology are found in Appendix A.

3 Results

3.1 Updating payoffs based on the last round with another member of the population

In this section we explore the case where the updating payoffs are based on the last round payoff achieved against another member of the population. We compare this to the expected payoffs. We assume that individuals interact in a donation game where each can cooperate by providing a benefit b to the other player at their cost c, with $0 < c < b$. Thus, $T = b, R = b - c, S = -c, P = 0$.

Figure 1 shows simulations results for the described process of section 2. Figure 1 depicts the evolving conditional cooperation probabilities p and q. The discount factor δ is comparably high, thus the opening move y is a transient effect and has no effect on the outcome. The left panel corresponds to the standard scenario considered in the literature. It considers players who use expected payoffs to update their strategies. The right panel shows the scenario considered herein, in which players update their strategies based on their last round's payoff. The top panels assume a benefit b of 3, whereas the bottom assume a benefit of 10.

The figure suggests that when updating is based on expected payoffs, players tend to be more generous and more cooperative. The q-values are higher on average which suggests that individuals tend to be more generous when it comes to cooperating after being at the receiving end of a defection. The average cooperation rate within the resident population is strictly higher when expected payoffs are considered. The difference between the two methods for both the values of b are statistically significant. This contrast becomes more obvious for $b = 10$. More specifically the average cooperation drops from 97% to 51%. The residents of the population cooperated on average 97% of the times in one case and only 57% in the other.

We further explore the effect of the benefit, Figure 2. The figure suggests that the expected payoffs always overestimate cooperation. For the limited memory payoffs, the cooperation rate remains unchanged at approximately 50% once $b = 5$. The highest cooperation rate that was achieved was 0.51 when the last round payoff is considered. In comparison, when the expected payoffs were used the highest average cooperation rate was 0.97.

Figure 3 illustrates results for various runs of the evolutionary process where we vary the strength of

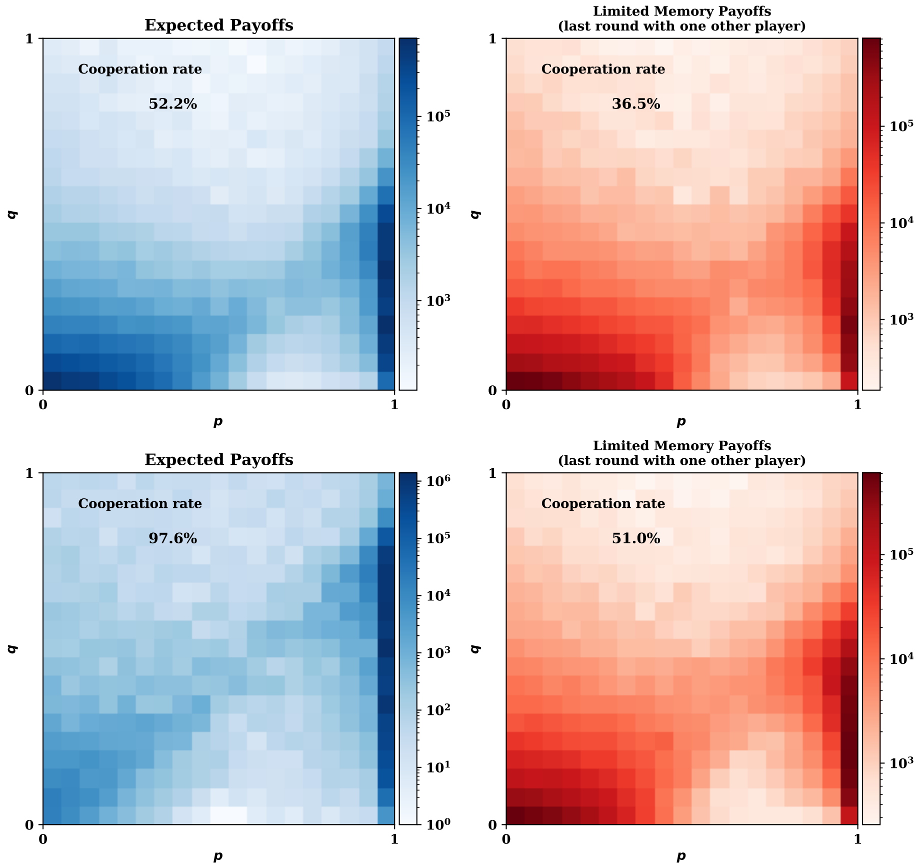


Figure 1: Evolutionary dynamics under expected payoffs and last round with one interaction payoffs. We have run two simulations of the evolutionary process described in section 2 for $T = 10^7$ time steps. For each time step, we have recorded the current resident population (y, p, q). Since simulations are run for a relatively high continuation probability of $\delta = 0.999$, we do not report the players' initial cooperation probability y. The graphs show how often the resident population chooses each combination (p, q) of conditional cooperation probabilities in the subsequent rounds. (**A**) If players update based on their expected payoffs, the resident population typically applies a strategy for which $p \approx 1$ and $q \leq 1 - c/b = 0.9$. (**B**) When players update their strategies based on their realized payoffs in the last round, there are two different predominant behaviors. The resident population either consists of defectors (with $p \approx q \approx 0$) or of conditional cooperators. In the latter case, the maximum level of q consistent with stable cooperation is somewhat smaller compared to the expected-payoff setting, $q < 0.5$. The cooperation rate within the resident population (averaged over all games and over all time steps) is close to 100%. Parameters: $N = 100$, $c = 1$, $\beta = 1$, $\delta = 0.999$.

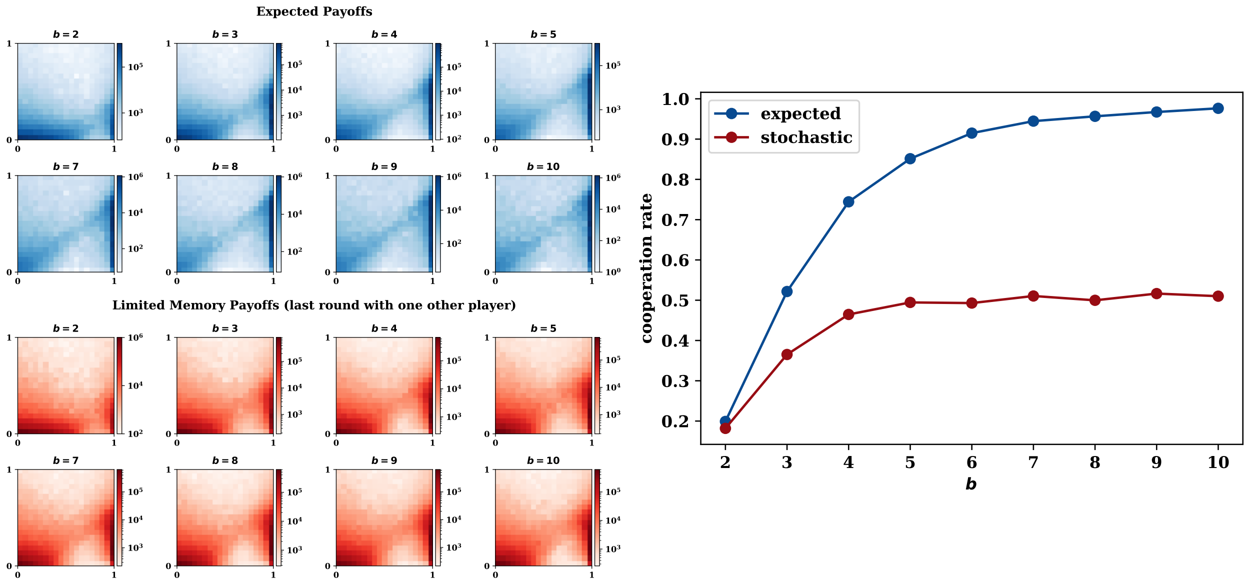


Figure 2: The evolution of cooperation for different benefit values. We vary the benefit of defection b. In all cases, expected payoffs appear to overestimate the average cooperation rate the population achieves. (**A**) the probabilities p, q for resident population over 10^7 time steps for each benefit value. (**B**) The cooperation rate within the resident population (averaged over all games and over all time steps) over the benefit. Unless explicitly varied, the parameters of the simulation are $N = 100$, $c = 1$, $\beta = 1$, $\delta = 0.99$. Simulations are run for $T = 5 \times 10^6$ time steps for each parameter combination.

selection. For weak selection, $\beta < 1$, the updating payoffs have no effect. The evolved population and the average cooperating rate are the same for both approaches. However for strong selection, it can be seen that the cooperating rate increases with $beta$, when expected payoffs are used. However, when limited memory payoffs are used the cooperating rate decreases.

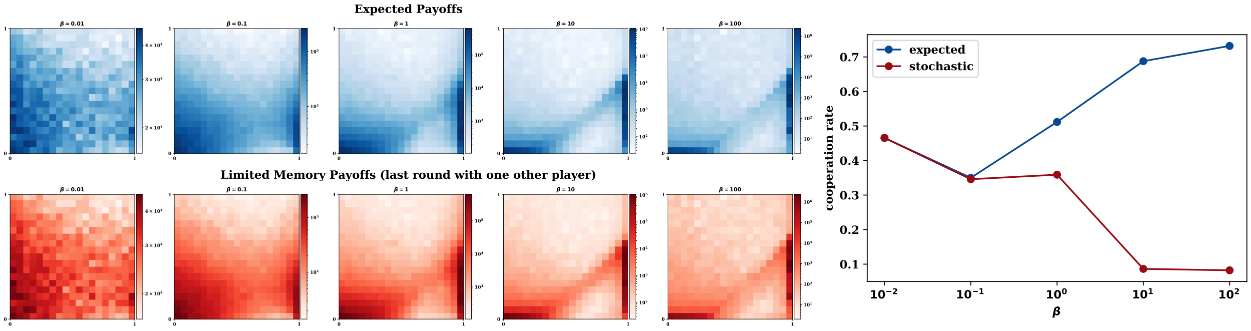


Figure 3: The evolution of cooperation for different selection strength values. We vary the selection strength β. In all cases, stochastic payoff evaluation tends to reduce the evolving cooperation rates. (**A**) the probabilities p, q for resident population over 10^7 time steps for each β value. (**B**) The cooperation rate within the resident population (averaged over all games and over all time steps) over β. Unless explicitly varied, the parameters of the simulation are $N = 100$, $b = 3$, $c = 1$, $\beta = 1$, $\delta = 0.99$. Simulations are run for $T = 5 \times 10^6$ time steps for each parameter combination.

3.2 Effect of updating payoffs in different social dilemmas

By focusing on the donation game in the previous section, we have gained insights into the possible effects of the updating payoffs, and how parameters such as the benefit and the strength of selection might can exaggerate the effect. However, by focusing on one specific game can narrow our knowledge of the effects of updating payoffs on the different forms of possible human interactions and social dilemmas. For this reason we explore not only the donation game but all the possible symmetric games. These are given by Table 1.

	social dilemmas	payoffs' constrains
(i)	stage hunt	$R > T > P > S$
(ii)	snowdrift	$T > R > S > P$
(iii)	harmony	$R > T > S > P$
(iv)	prisoner dilemma	$T > R > P > S$
(v)	donation game	$T > R > P > S$ & $T = b, R = b - c, S = -c, P = 0$

Table 1: Social dilemmas and payoffs' constrains. The various social dilemmas we explore in this work. Results for cases (i) - (iv) are presented in section 3.2 and results for case (v) are presented in section 3.1.

Without loss of generality we set one of the entries of the payoff matrix to 0 and another one to 1. A common normalization is $R = 1$ and $P = 0$ [6, 7]. We vary the temptation to defect T and the risk of cooperating S. We continue the discussion of the previous section. We compare the expected payoffs to the last round payoffs over the different social dilemmas. The results of the simulation runs are given in Figures 4 and 5.

At the left up quadrat ($S > 0$ and $T \leq 1$) the individuals interact in the harmony game at each game stage. The harmony game is a game without tensions. The best play in the harmony game is to always cooperate. Indeed both figures confirm this. The behaviour varies sightly at the case $T = 1$. Now the temptation is equal to the payoff of a cooperation and thus individuals now tend to defect but only slightly. The cooperating rate still remains high. At the bottom left up quadrat ($S \leq 0$ and $T < 1$) the individuals interact in the stag hunt game at each game stage. In the stag hunt game cooperation is still favoured. The best response of individuals is to cooperate and this is again confirmed by both approaches.

The last two cases are those of the prisoner's dilemma and of the snowdrift game. At the top right quadrat we explore the snowdrift game. In the snowdrift game individuals want to cooperate after receiving a defection and to defect given that their opponent defected. This behavior is captured by the figures. In the case of q, cooperating after receiving a cooperation, we observe a decrease as the suckers payoff decreases. This is to be expected. Finally, in the right bottom quadrat individuals interact in a prisoner's dilemma. In the prisoner's dilemma individuals tend to defect. In the case of direct reciprocity cooperation can emerge, as we discussed in the previous section. However, in the spectrum of all the different values of S and T used here we observe this behaviour in only a handful of cases.

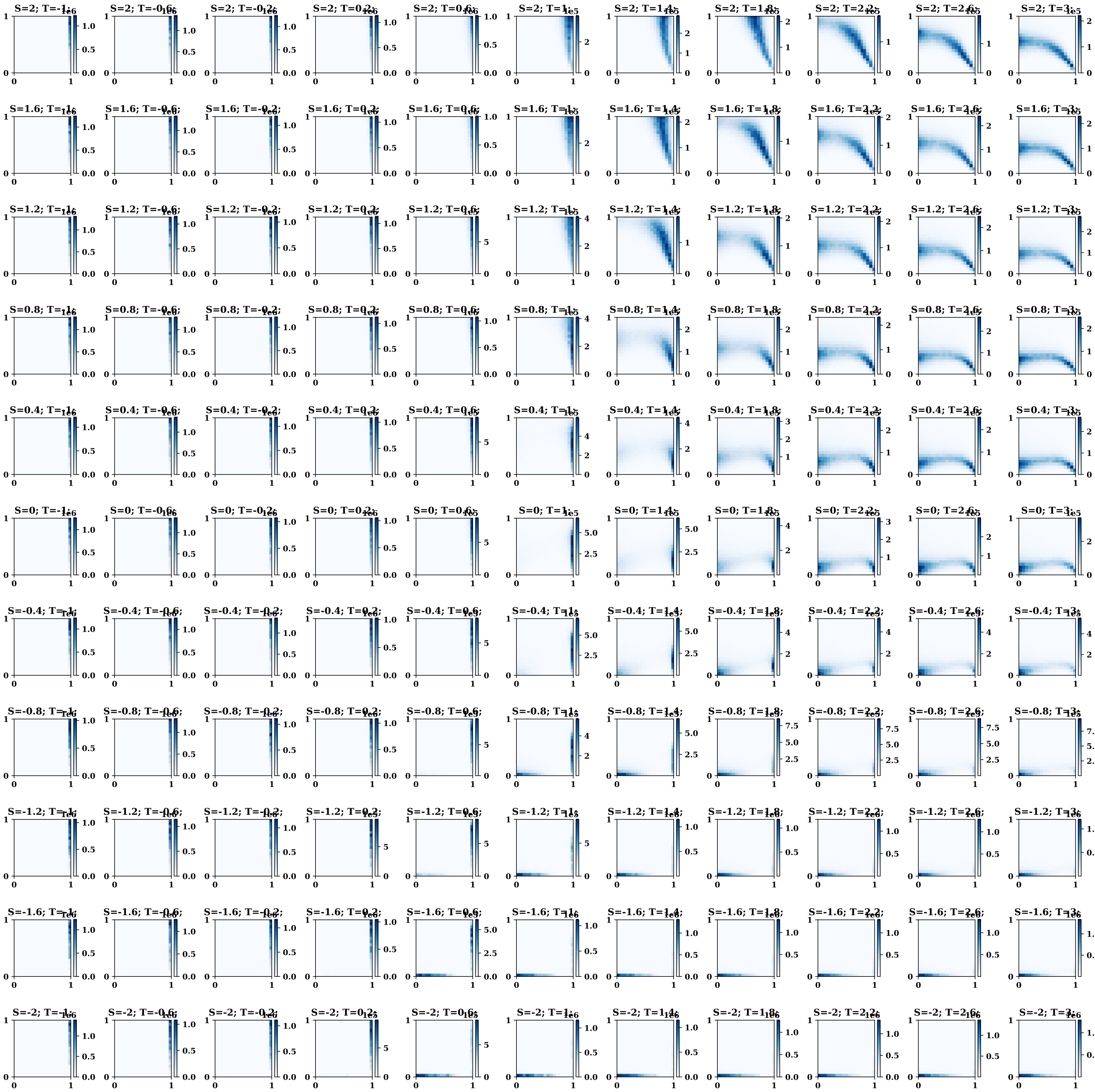


Figure 4: Evolutionary dynamics under expected payoffs for various social dilemmas. We have run several simulations of the evolutionary process described in section 2 for $T = 10^7$ time steps. The graphs show how often the resident population chooses each combination (p, q) of conditional cooperation probabilities in the subsequent rounds. We vary the temptation payoff $T \in \{-1, -0.6, -0.2, 0.2, 0.6, 1, 1.4, 1.8, 2.2, 2.6, 3\}$ across the x axis, and $S \in \{2, 1.6, 1.2, 0.8, 0.4, 0, -0.4, -0.8, -1.2, -1.6, -2\}$ across the y axis. Parameters: $N = 100$, $c = 1$, $\beta = 1$, $\delta = 0.999$.

We have discussed the expected behaviour for each social dilemma. We now want to account the difference in the cooperation rate given the two updating payoffs. The cooperation rates for each of the case of the

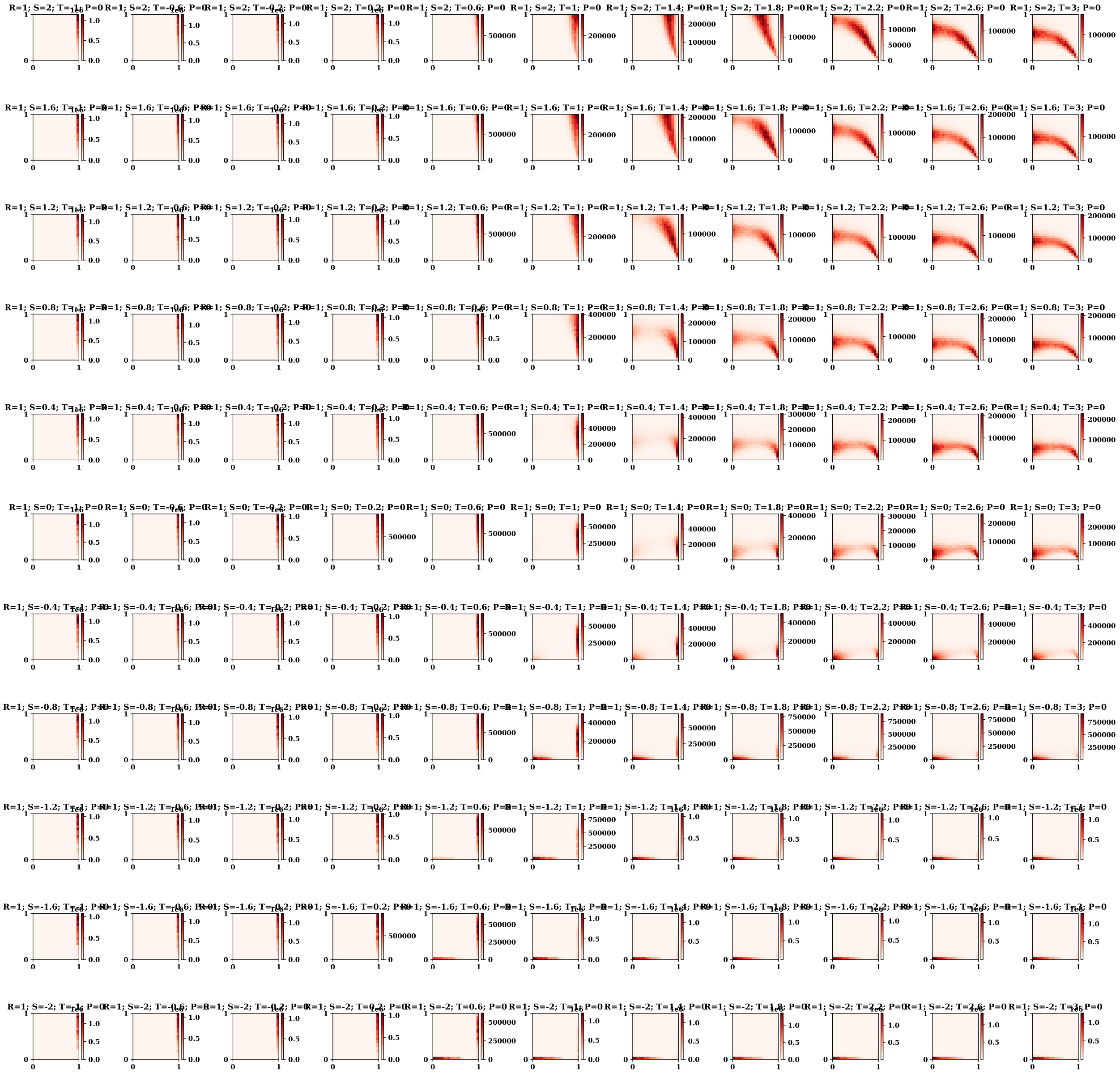


Figure 5: Evolutionary dynamics under last round payoffs for various social dilemmas. We have run several simulations of the evolutionary process described in section 2 for $T = 10^7$ time steps. The graphs show how often the resident population chooses each combination (p, q) of conditional cooperation probabilities in the subsequent rounds. We vary the temptation payoff $T \in \{-1, -0.6, -0.2, 0.2, 0.6, 1, 1.4, 1.8, 2.2, 2.6, 3\}$ across the x axis, and $S \in \{2, 1.6, 1.2, 0.8, 0.4, 0, -0.4, -0.8, -1.2, -1.6, -2\}$ across the y axis. Parameters: $N = 100$, $c = 1$, $\beta = 1$, $\delta = 0.999$.

social dilemmas we have used are given in Figure 6.

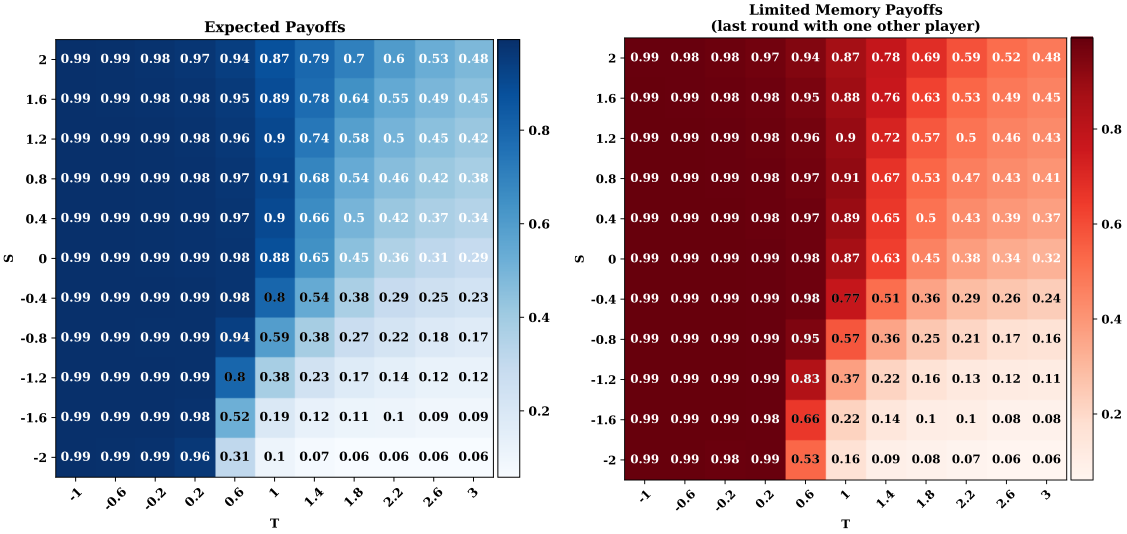


Figure 6: Cooperation rates over various social dilemmas. (**A**) If players update based on their expected payoffs. (**B**) If players update based on their last round payoffs.

So far we have explored the difference between the expected payoffs and the last round payoffs. In order to explore further the effect of limited memory we allow individuals to remember more. We consider the cases where individuals remember up two interactions, and up to the last two rounds. In total we present results for three more cases. These are the cases of the last round two rounds with another member of the population, last round with two members of the population, and last round two rounds with two members of the population.

Similarly to the previous examples we have run the evolutionary process for a large number of step for each of the social dilemmas. The results remain the same. The behavior over the different games remains the same. We note that now there is more noise in the evolved populations. Due to space the figures have been moved to the Appendix, however, the cooperating rates for each of the cases are given in Figure 7.

4 Conclusions

A Model Setup

Consider a population of N individuals where N is even. At any point in time there are at most two different strategies in present in the population. More specifically, a mutant strategy played by k individuals and a resident strategy played by $N - k$ individuals. We assume a pairwise process in which strategies spread because they are imitated more often. Each step of the evolutionary process consists of two stages; a game stage and an update stage.

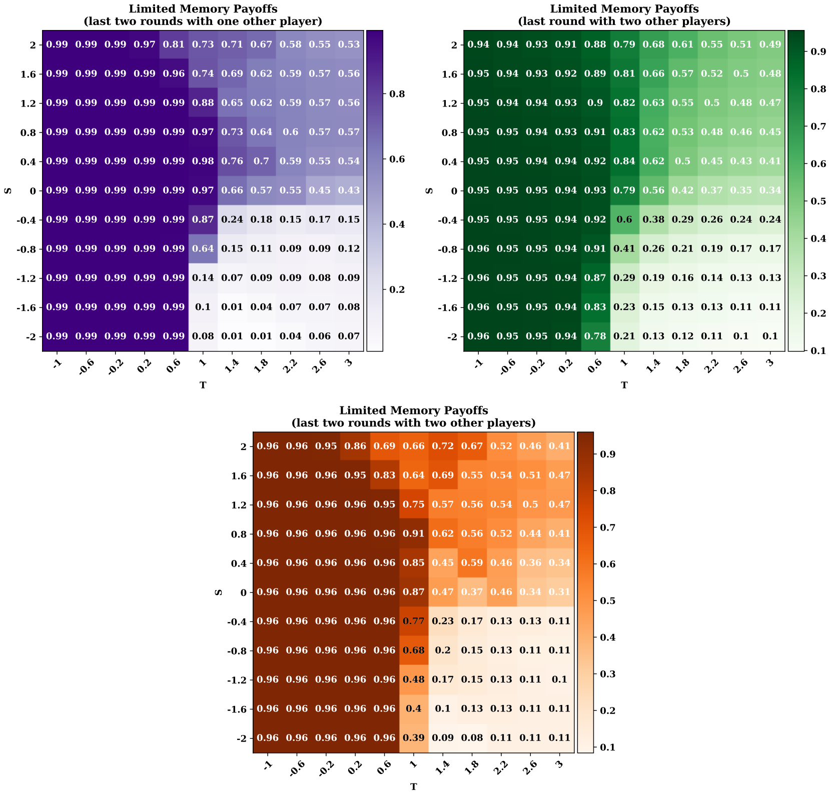


Figure 7: Cooperation rates over various social dilemmas for different limited memory payoffs. (**A**) If players update based on their last two rounds with another member of the population. (**B**) If players update based on their last round with two other members of the population. (**C**) If players update based on their last two rounds with two other members of the population.

In the game stage, each individual is randomly matched with some other individual in the population. Their interaction lasts for a number of turns which is not fixed but depends on the continuation probability δ. At each turn the individuals choose between cooperation (C) and defection (D). Thus, there are four possible outcomes in each turn CC, CD, DC and DD. If both players cooperate they receive the reward payoff R, whereas if both players defect they receive the punishment payoff P. If one cooperates but the other defects, the defector receives the temptation to defect, T, whereas the cooperator receives the sucker's payoff, S. Let $\mathcal{U} = \{R, S, T, P\}$ denote the set of feasible payoffs in each round, and let $\mathbf{u} = (R, S, T, P)$ be the corresponding payoff vector. The values of the payoffs are not only based on the prisoner's dilemma but all the symmetric 2×2 games, Table 1.

A further assumption of our model is that individuals make use of reactive strategies when they make decisions in each round. Reactive strategies are a set of strategies that take into account only the previous action of the opponent. A reactive strategy can be written explicitly as a vector,

$$s = (y, p, q)$$

where y is the probability that the strategy opens with a cooperation and p, q are the probabilities that the strategy cooperates given that the opponent cooperated and defected equivalently.

In the updating stage, two players are randomly drawn from the population, a 'learner' and a 'role model'. The learner adopts the role model's strategy based on the Fermi distribution function,

$$\rho(u_L, u_{RM}) = \frac{1}{1+\exp^{-\beta(u_{RM}-u_L)}}. \tag{2}$$

where $u_L \in \mathcal{U}$ is the learner's payoff, $u_{RM} \in \mathcal{U}$ is the role model's payoff, and $\beta \geq 0$ is the intensity of selection.

We iterate this basic evolutionary step until either the mutant strategy goes extinct, or until it fixes in the population and becomes the new resident strategy. After either outcome, we set k to 1 and we introduce a new mutant strategy which is uniformly chosen from all reactive strategies at random. Instead of simulating each step of the evolutionary process, we estimate the probability that a newly introduced mutant fixes [8]. This is defined as the fixation probability of the mutant, and the standard form is the following,

$$\varphi = \frac{1}{1 + \sum\limits_{i=1}^{N-1} \prod\limits_{k}^{i} \frac{\lambda_k^-}{\lambda_k^+}}, \tag{3}$$

where λ_k^-, λ_k^+ are the probabilities that the number of mutants decreases and increases respectively.

This process of mutation and fixation/extinction is iterated many times. The evolutionary process is summarized by Algorithm 1.

Algorithm 1: Evolutionary process

```
$N \leftarrow$ population size;
$k \leftarrow 1$;
resident $\leftarrow (0, 0, 0)$;
while step $<$ maximum number of steps do
    mutant $\leftarrow$ random: $\{\emptyset\} \rightarrow R^3$;
    fixation probability $\leftarrow \varphi$;
    if $\varphi >$ random: $i \rightarrow [0, 1]$ then
        resident $\leftarrow$ mutant;
    end
end
```

The aim of this work is to explore the effect of updating memory on the cooperation rate of the evolved population. For this reason we consider two different approaches when estimating the payoffs at the updating stage. The two approaches we consider are those of (i) the expected and (ii) the limited memory payoffs.

Expected Payoffs

The expected payoffs are the conventional payoffs used in the updating stage [9]. They are defined as the mean payoff of an individual in a well-mixed population that engages in repeated games with all other population members.

We first define the payoff of two reactive strategies at the game stage. Assume two reactive strategies $s_1 = (y_1, p_1, q_1)$ and $s_2 = (y_2, p_2, q_2)$. It is not necessary to simulate the play move by move, instead the play between the two strategies is defined a Markov matrix M,

$$M = \begin{bmatrix} p_1p_2 & p_1\left(1-p_2\right) & p_2\left(1-p_1\right) & \left(1-p_1\right)\left(1-p_2\right) \\ p_2q_1 & q_1\left(1-p_2\right) & p_2\left(1-q_1\right) & \left(1-p_2\right)\left(1-q_1\right) \\ p_1q_2 & p_1\left(1-q_2\right) & q_2\left(1-p_1\right) & \left(1-p_1\right)\left(1-q_2\right) \\ q_1q_2 & q_1\left(1-q_2\right) & q_2\left(1-q_1\right) & \left(1-q_1\right)\left(1-q_2\right) \end{bmatrix}. \tag{4}$$

whose stationary vector $\mathbf{v}$, combined with the payoff u, yields the game stage outcome for each strategy, $\langle \mathbf{v}(s_1, s_2), \mathbf{u} \rangle$ [3].

In the updating stage the learner adopts the strategy of the role model based on their updating payoffs. Given that there are only two different types in the population at each time step we only need to define the expected payoff for a resident (π_R) and for a mutant (π_M). Assume the resident strategy $s_R = (y_R, p_R, q_R)$ and the mutant strategy $s_M = (y_M, p_M, q_M)$, the expected payoffs are give by,

$$
\begin{aligned}
\pi_R \quad &= \quad \frac{N-k-1}{N-1} \cdot \langle \mathbf{v}(s_R, s_R), \mathbf{u} \rangle \quad + \quad \frac{k}{N-1} \cdot \langle \mathbf{v}(s_R, s_M), \mathbf{u} \rangle, \\
\pi_M \quad &= \quad \frac{N-k}{N-1} \cdot \langle \mathbf{v}(s_M, s_R), \mathbf{u} \rangle \quad + \quad \frac{k-1}{N-1} \cdot \langle \mathbf{v}(s_M, s_M), \mathbf{u} \rangle.
\end{aligned}
\tag{5}
$$

The number of mutant in the population increase if a learner resident adopts the strategy of a mutant role model, and decreases if a mutant leaner adopts the strategy of a resident. The probabilities that the number of mutants decreases and increases, λ_k^- and λ_k^+, are not explicitly define as,

$$
\begin{aligned}
\lambda_k^- &= \rho(\pi_R, \pi_M) \\
\lambda_k^+ &= \rho(\pi_M, \pi_R).
\end{aligned}
$$

Limited memory payoffs

Initially, we discuss the case of the last round updating payoff. At the stage game we define the payoff of a reactive strategy in the last round, Proposition **??**.

Proposition 1. *Consider a repeated prisoner's dilemma, with continuation probability δ, between players with reactive strategies $s_1 = (y_1, p_1, q_1)$ and $s_2 = (y_2, p_2, q_2)$ respectively. Then the probability that the s_1 player receives the payoff $u \in \mathcal{U}$ in the very last round of the game is given by $v_u(s_1, s_2)$, as given by Equation (6).*

$$
\begin{aligned}
v_R(s_1, s_2) &= (1-\delta)\frac{y_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\Big(q_1 + r_1\big((1-\delta)y_2 + \delta q_2\big)\Big)\Big(q_2 + r_2\big((1-\delta)y_1 + \delta q_1\big)\Big)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)} \times R, \\
v_S(s_1, s_2) &= (1-\delta)\frac{y_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\Big(q_1 + r_1\big((1-\delta)y_2 + \delta q_2\big)\Big)\Big(\bar{q}_2 + \bar{r}_2\big((1-\delta)y_1 + \delta p_1\big)\Big)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)} \times S, \\
v_T(s_1, s_2) &= (1-\delta)\frac{\bar{y}_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\Big(\bar{q}_1 + \bar{r}_1\big((1-\delta)y_2 + \delta p_2\big)\Big)\Big(q_2 + r_2\big((1-\delta)y_1 + \delta q_1\big)\Big)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)} \times T, \\
v_P(s_1, s_2) &= (1-\delta)\frac{\bar{y}_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\Big(\bar{q}_1 + \bar{r}_1\big((1-\delta)y_2 + \delta p_2\big)\Big)\Big(\bar{q}_2 + \bar{r}_2\big((1-\delta)y_1 + \delta p_1\big)\Big)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)} \times P.
\end{aligned}
\tag{6}
$$

In these expressions, we have used the notation $r_i := p_i – q_i$, $\bar{y}_i = 1 – y_i$, $\bar{q}_i := 1 – q_i$, and $\bar{r}_i := \bar{p}_i – \bar{q}_i = -r_i$

for $i \in \{1, 2\}$.

Proof. Given a play between two reactive strategies with continuation probability δ. The outcome at turn t is given by,

$$(1-\delta)\mathbf{v_0}\sum \delta^t M^{(t)}, \tag{7}$$

where $\mathbf{v_0}$ denotes the expected distribution of the four outcomes in the very first round, and $1-\delta$ the probability that the game ends. It can be shown that,

$$\begin{aligned}
(1-\delta)\mathbf{v_0}\sum \delta^t M^{(t)} &= (1-\delta)(\mathbf{v_0} + \delta \mathbf{v_0} M + \delta^2 \mathbf{v_0} M^2 + \dots) \\
&= (1-\delta)\mathbf{v_0}(1 + \delta M + \delta^2 M^2 + \dots) \text{ using standard formula for geometric series} \\
&= (1-\delta)\mathbf{v_0}(I_4 - \delta M)^{-1}
\end{aligned}$$

where $(1-\delta)\mathbf{v_0}(I_4 - \delta M)^{-1}$ is vector $\in R^4$ and it the probabilities for being in any of the outcomes CC, CD, DC, DD in the last round. Combining this with the payoff vector u and some algebraic manipulation we derive to the Equation 6. □

In the updating stage we select a mutant and resident to be either the role model or the learner. Assume the selected mutant. Given that they can interact with only one other member of the population, they can interact either the selected resident, another resident or with another mutant. The same is true for the the selected resident. They can interact with the selected mutant, another resident, or another mutant. Thus, in each updating stage there are five possible combinations of pairs. These are illustrated by Figure 8.

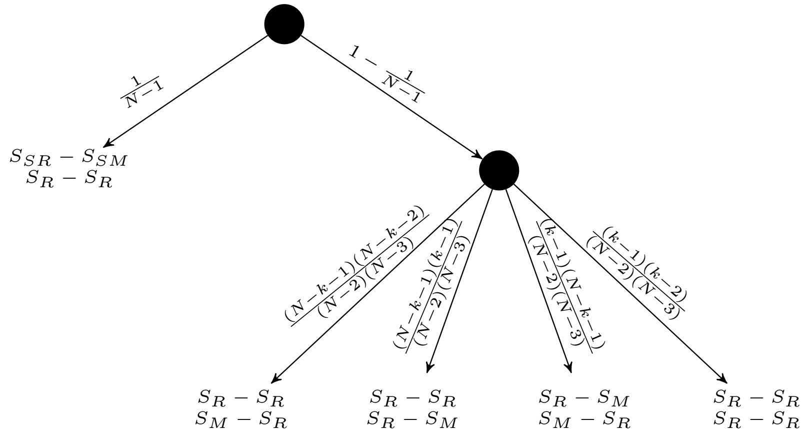


Figure 8: Possible pairings combination in the updating stage, given that individuals interact with only one other member in the population.. We distinguish between the selected resident and the rest of the residents and we do the same with the mutants. There is a probability that the selected resident interacts with the selected mutant.

The probability that the respective payoffs of the players are given by u_1 and u_2 can be calculated as

$$
\begin{aligned}
x(u_1,u_2) = \qquad & \frac{1}{N-1}\cdot v_{u_1}(S_1,S_2)\cdot 1_{(u_1,u_2)\in\mathcal{U}_F^2} \\
&+\left(1-\frac{1}{N-1}\right)\Big[\tfrac{k-1}{N-2}\,\tfrac{k-2}{N-3}v_{u_1}(S_1,S_2)v_{u_2}(S_2,S_2)+\tfrac{k-1}{N-2}\,\tfrac{N-k-1}{N-3}v_{u_1}(S_1,S_2)v_{u_2}(S_2,S_1) \\
&\qquad +\tfrac{N-k-1}{N-2}\,\tfrac{k-1}{N-3}v_{u_1}(S_1,S_1)v_{u_2}(S_2,S_2)+\tfrac{N-k-1}{N-2}\,\tfrac{N-k-2}{N-3}v_{u_1}(S_1,S_1)v_{u_2}(S_2,S_1)\Big]\,.
\end{aligned}
\tag{8}
$$

The first term on the right side corresponds to the case that the learner and the role model happened to be matched during the game stage, which happens with probability $1/(N-1)$.In that case, we note that only those payoff pairs can occur that are feasible in a direct interaction, $(u_1,\, u_2)\in\mathcal{U}_F^2 := \big\{(R,\, R),\, (S, T),\, (T,$ $S),\, (P, P)\big\}$, as represented by the respective indicator function. Otherwise, if the learner and the role model did not interact directly, we need to distinguish four different cases, depending on whether the learner was matched with a resident or a mutant, and depending on whether the role model was matched with a resident or a mutant.

Given that $N-k$ players use the resident strategy $S_1=(y_1,p_1,q_1)$ and that the remaining k players use the mutant strategy $S_2=(y_2,p_2,q_2)$, the probability that the number of mutants increases by one in one step of the evolutionary process can be written as

$$
\lambda_k^+ = \frac{N-k}{N}\cdot\frac{k}{N}\cdot\sum_{u_1,u_2\in\mathcal{U}} x(u_1,u_2)\cdot\rho(u_1,u_2), \tag{9}
$$

$$
\lambda_k^- = \frac{N-k}{N}\cdot\frac{k}{N}\cdot\sum_{u_1,u_2\in\mathcal{U}} x(u_1,u_2)\cdot\rho(u_2,u_1). \tag{10}
$$

In this expression, $(N-k)/N$ is the probability that the randomly chosen learner is a resident, and k/N is the probability that the role model is a mutant. The sum corresponds to the total probability that the learner adopts the role model's strategy over all possible payoffs u_1 and u_2 that the two player may have received in their respective last rounds. We use $x(u_1,u_2)$ to denote the probability that the randomly chosen resident obtained a payoff of u_1 in the last round of his respective game, and that the mutant obtained a payoff of u_2.

This framework can be extended to consider the case of where the payoffs correspond to the last n rounds payoff an individual achieved after interacting with m other individuals. For the case $n = 2$ the payoffs at the game stage are,

Proposition 2. *Assume a play between the reactive strategies s_1 and s_2 with a continuation probability δ. Then the probability of being in any of the sixteen outcomes RR, RR, RR, RR, RR, RR, RR, RR, RR,*

$RR, RR, RR, RR, RR, RR, RR$ on the last two rounds are given by,

$$\mathbf{v_{a_1,a_2}} = (1-\delta) m_{a_1,a_2} \delta^2 \left[\mathbf{v_0}(I_4 - \delta M)^{-1}\right]_{a_1,a_2}, \quad \textit{for } m_{a_1,a_2} \in M \;\&\; a_1, a_2 \in \{R, S, T, P\} \tag{11}$$

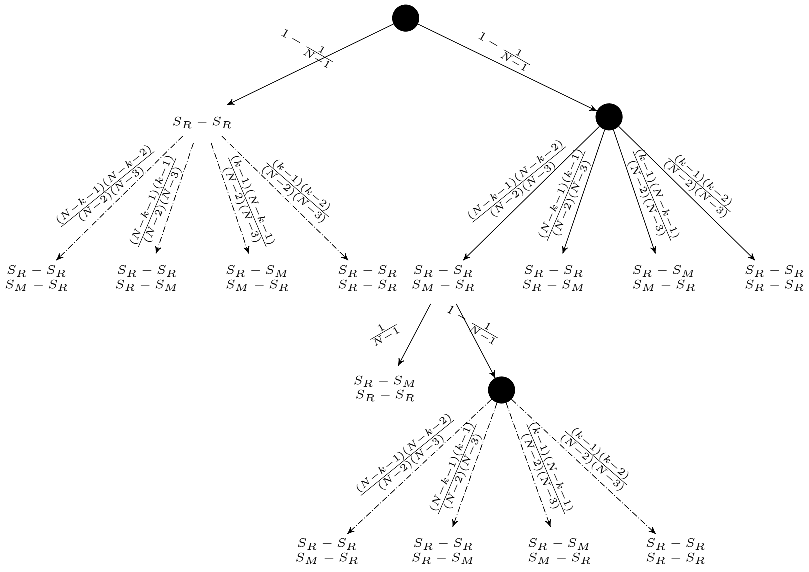
Proposition 2 can be extended to the last n rounds.

Proposition 3. *Assume a play between the reactive strategies s_1 and s_2 with a continuation probability δ. Then the probability of being in any of the sixteen outcomes RR, RR, RR, RR, RR, RR, RR, RR, RR, $RR, RR, RR, RR, RR, RR, RR$ on the last two rounds are given by,*

$$\mathbf{v_{a_1,a_2}} = (1-\delta) \prod m_{a_1,a_2} \delta^2 \left[\mathbf{v_0}(I_4 - \delta M)^{-1}\right]_{a_1,a_2} \tag{12}$$

for $m_{a_1,a_2} \in M$ and $a_1, a_2 \in [1, 4]$.

Equation 8 can also be extended to include interactions with two other individuals. The possible pairings are illustrated by Figure **??**.



$1 - \frac{1}{N-1}$
$1 - \frac{1}{N-1}$
$S_R - S_R$
$\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(N-k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(N-k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(k-2)}{(N-2)(N-3)}$
$S_R - S_R$ $S_M - S_R$
$S_R - S_R$ $S_R - S_M$
$S_R - S_M$ $S_M - S_R$
$S_R - S_R$ $S_R - S_R$
$S_R - S_R$ $S_M - S_R$
$S_R - S_R$ $S_R - S_M$
$S_R - S_M$ $S_M - S_R$
$S_R - S_R$ $S_R - S_R$
$\frac{1}{N-1}$
$1 - \frac{1}{N-1}$
$S_R - S_M$ $S_R - S_R$
$\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(N-k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(k-2)}{(N-2)(N-3)}$
$S_R - S_R$ $S_M - S_R$
$S_R - S_R$ $S_R - S_M$
$S_R - S_M$ $S_M - S_R$
$S_R - S_R$ $S_R - S_R$

Figure 9: The tree

B Verifying analytical results with simulations

The analytical results presented in this work have been verified with simulations. More specifically the probabilities of Equation (6),

Proposition 2,

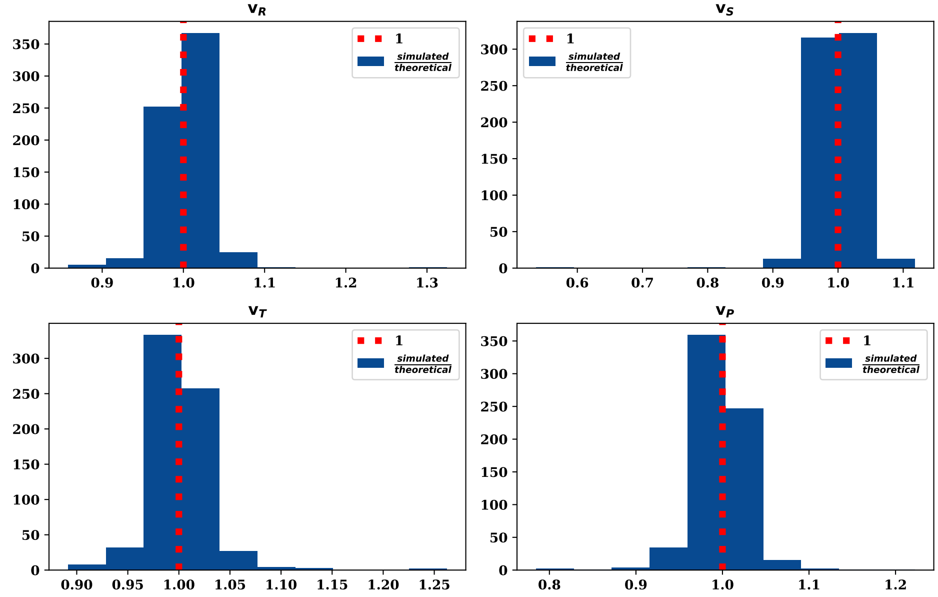


Figure 10

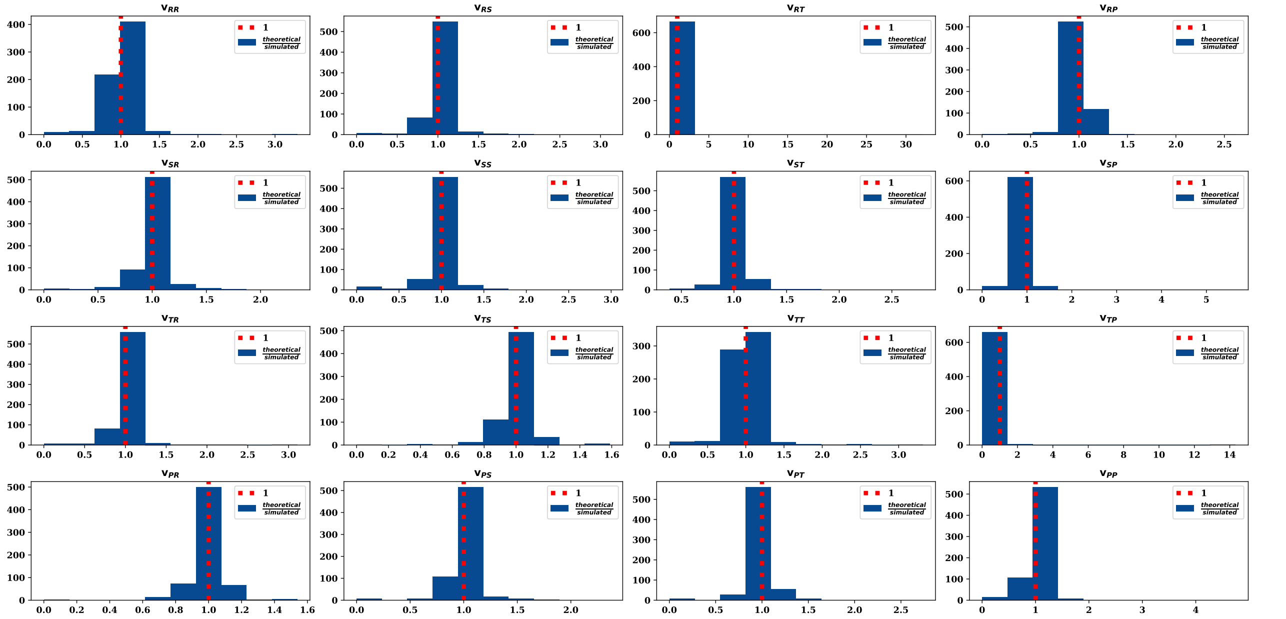


Figure 11

References

[1] Martin A Nowak and Karl Sigmund. Tit for tat in heterogeneous populations. *Nature*, 355(6357):250–253, 1992.

[2] Seung Ki Baek, Hyeong-Chai Jeong, Christian Hilbe, and Martin A Nowak. Comparing reactive and memory-one strategies of direct reciprocity. *Scientific reports*, 6(1):1–13, 2016.

[3] Ch Hauert and Heinz Georg Schuster. Effects of increasing the number of players and memory size in the iterated prisoner's dilemma: a numerical approach. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1381):513–519, 1997.

[4] Alexander J Stewart and Joshua B Plotkin. Small groups and long memories promote cooperation. *Scientific reports*, 6(1):1–11, 2016.

[5] Carlos P. Roca, José A. Cuesta, and Angel Sánchez. Time scales in evolutionary dynamics. *Phys. Rev. Lett.*, 97:158701, Oct 2006.

[6] Luis A Martinez-Vaquero, Jose A Cuesta, and Angel Sanchez. Generosity pays in the presence of direct reciprocity: A comprehensive study of 2× 2 repeated games. *PLoS One*, 7(4):e35135, 2012.

[7] Carlos P Roca, José A Cuesta, and Angel Sánchez. Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. *Physics of life reviews*, 6(4):208–249, 2009.

[8] Martin A Nowak, Akira Sasaki, Christine Taylor, and Drew Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428(6983):646–650, 2004.

[9] Lorens A Imhof and Martin A Nowak. Stochastic evolutionary dynamics of direct reciprocity. *Proceedings of the Royal Society B: Biological Sciences*, 277(1680):463–468, 2010.