Evolution of cooperation among individuals with limited payoff memory

Nikoleta E. Glynatsi, Christian Hilbe, Alex McAvoy

Abstract

1 Introduction

Evolutionary game theory [1–4] describes the evolutionary dynamics of populations consisting of different types of interacting individuals. Traditional approaches assume that individuals meet each other at random in an infinitely large well-mixed population. The replicator dynamics describes how the abundance of strategic types in a population changes based on their fitness, identified with the payoff resulting from the game [5]. In this deterministic formulation, individuals with higher fitness, such as defectors, increase in abundance. Ultimately, the system reaches a stable fixed point in which the population may consist either of a single type or of a mixture of different types.

It has been shown that the finiteness of populations may lead to fundamental changes in this picture due to stochastic effects In this framework disadvantageous mutants have a small, yet non-zero probability to reach fixation in a finite population.

Its most prominent application is the quest for the origins and evolution of cooperation. An important application of evolutionary game theory is the evolution of cooperation which remains one of the greatest problems for biological and social sciences.

, and in general, defectors are stable against invasion by cooperators.

In this framework disadvantageous mutants have a small, yet non-zero probability to reach fixation in a finite population.

Two classes of microscopic processes have been used extensively: (i) fitness-based processes in which an individual chosen proportional to fitness reproduces and the offspring replaces a randomly chosen individual [11]; (ii) pairwise comparison processes in which a pair of individuals is chosen, and where subsequently one of these individuals may adopt the strategy of the other.

In the pairwise process considers a finite population of fixed size.In the population at a time step different types of individuals can exist. Each individual can be of one of two types, A and B. The state of the population is thus characterized by the number i of individuals of type A.

The interaction between the two types of individuals is described by the functions π_{A}^{i} and π_{B}^{i}. These indicate the expected payoff for two types in a population in state i. The interaction can be generated from a two-player matrix game which leads to payoffs linear in i [26], but we keep the formalism general to include games played between an arbitrary number of players, which leads to payoffs that depend on i in a polynomial way [27–30]. In fact, our results hold for an arbitrary dependence of the payoffs on i.

At each step an individual change their type based. This very cooperators can repeas.

This model has several assumptions. Two assumptions are that initally it is assumed that time to interact is faster than the time of the population update. Secondly it is assumed that the indiviail can interact will every other population member, and recall their interaction perfecttly.

The first assumption was questioned by the work of.

In this work

We first consider two extreme scenarios, the classical scenario and the alternative scenario where individuals update their strategies only based on the very last payoff they obtained. We observe that individuals with limited memory tend to adopt less generous strategies and they achieve less cooperation when interacting in a prisoner's dilemma. We obtain similar results when we consider that individuals update their strategies based on more information. More specifically, up to the last two payoffs they obtained when interacting with up to two different members of the population. We extend our approach to the rest of the symmetric 2 × 2 games.

The remainder of the paper is organized as follows. In section 2 we describe the model. In section 3 we present the results of the simulations, and in section 4 we outline the main conclusions.

2 Model Setup

In the following, we consider a well mixed population of fixed size. In each step of t

We study the transmission of strategies with a frequency-dependent birth-death process26 in a finite population of size n. In each time step, two randomly chosen individuals compare their payoffs and one of them can switch to the other one's strategy. This process can be interpreted as a model for social learning, whereby successful strategies spread, and, occasionally, random strategy exploration introduces novel strategies (corresponding to mutations in biological models).

We consider a population of N players [1] where N is even and mutations are sufficiently rare. Therefore, at any point in time there are at most two different strategies present in the population; a *resident* strategy and a *mutant* strategy. To describe how strategies spread we use a pairwise comparison process [6]. Each step of the evolutionary process consists of two stages, a game stage and an updating stage.

[1] The terms "player" and "individual" are used interchangeably here.

In the game stage each individual is randomly matched with some other individual in the population. They engage in a match where each subsequent turn occurs with a fixed probability δ. At each turn players choose independently to either cooperate (C) or to defect (D), and the payoffs of the turn depend on both their decisions. If both players 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 payoff T, whereas the cooperator receives the sucker's payoff S. We denote the feasible payoff of each turn as $\mathcal{U} = \{R, S, T, P\}$. We assume 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 $s = (y, p, q)$. The parameter 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 the learner's payoff $u_L \in \mathcal{U}$ and the role model's payoff $u_{RL} \in \mathcal{U}$, the learner adopts the role model's strategy with probability,

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

where $\beta \geq 0$ is the strength of selection. For small values of β the imitation probability is independent of the strategies of the involved players. As the value of β increases, the more likely it is that the learner adopts only strategies that yield a higher payoff. Conventionally the updating payoffs of the learner and the role model are based on their expected payoffs. A player's expected payoff is the mean payoff the player yields after engaging in matches of multiple turns with each member of the population. At each match a player bases their next turn decision only on the previous action of the opponent, however, the same player bases their expected payoffs on the outcomes of all their matches. Thus, a player is assumed to have limited and perfect memory at the same time. We propose a new a set of updating payoffs where it is also assumed that a player has limited memory. We referee to these as the limited memory payoffs.

The 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.

In order to account for the effect of the updating payoffs we simulate the evolutionary process and record which strategies the players adopt over time based on (i) the expected payoffs (ii) the limited memory payoffs. We compare the difference in the cooperation rate within the resident population for the two approaches. To account for the various types of social behaviour we also present results on multiple social dilemmas.

3 Results

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

To investigate the role of extortion in the context of evolutionary games, we concentrate on the donation game

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 and we compare this to the classical scenario of the expected payoffs. We assume that each pair of players interacts in a donation game. The donation game is a special case of the prisoner's dilemma. Each player can choose to cooperate by providing a benefit b to the other player at their cost c, with $0 < c < b$. Thus, the feasible payoffs in each round are $\mathcal{U} = \{b - c, -c, b, 0\}$.

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 we do not report the opening move y as it is a transient effect. The left panels correspond 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 of the resident strategies are on average higher in the case of the expected payoffs. The players will occasionally forgive a defection more often if their fitness depends on interacting with every member of the population. On the other hand, when social interactions are limited they are less forgiving. The average cooperation rate for each simulation is calculated as the average cooperation rate within the resident population. In the case of the expected payoffs, regardless the value of benefit, the average cooperation rate is strictly higher than that of the last round payoffs. The difference based on the two methods is statistically significant, and in the case of $b = 10$ the average cooperation of resident strategies drops from 97% to 57%.

We further explore the effect of benefit in Figure 2. The figure suggests that expected payoffs always yield a higher cooperation rate. In the case of expected payoffs we observe that the cooperation rate increases as the value of the benefit gets higher. In comparison for the limited memory payoffs, the cooperation rate remains unchanged at approximately 50% once $b = 5$.

We also investigate the effect of the strength of selection β. Figure 3 illustrates results for various runs of the evolutionary process. For weak selection, $\beta < 1$, we observe that the two methods yield similar results, however, as β increases there is variation in the evolving populations. In the case of expected payoffs the resident populations become more cooperative as β increases, whereas in the case of limited memory payoffs, the resident populations become more defective.

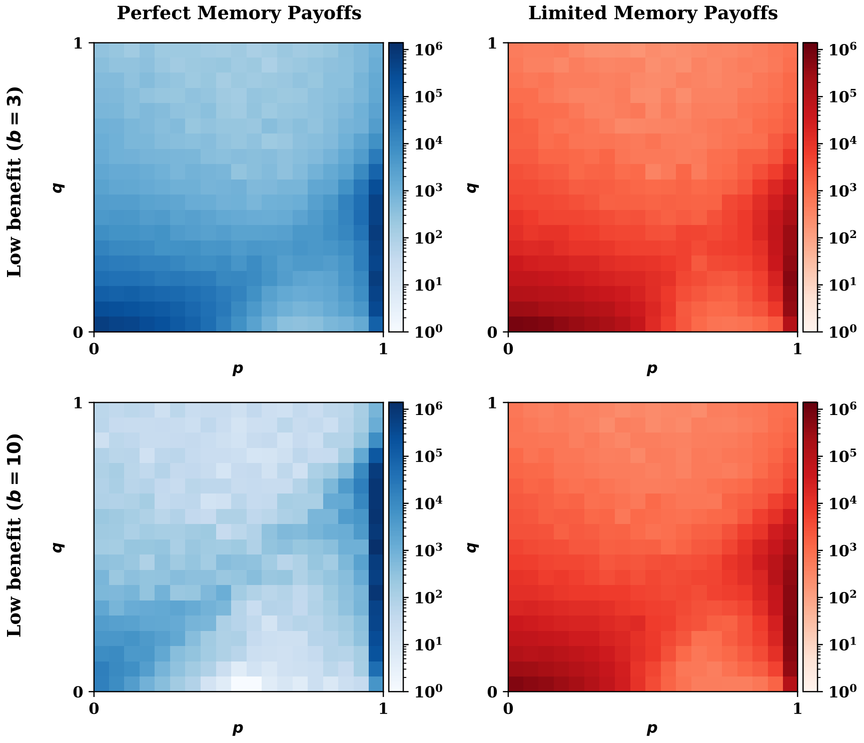
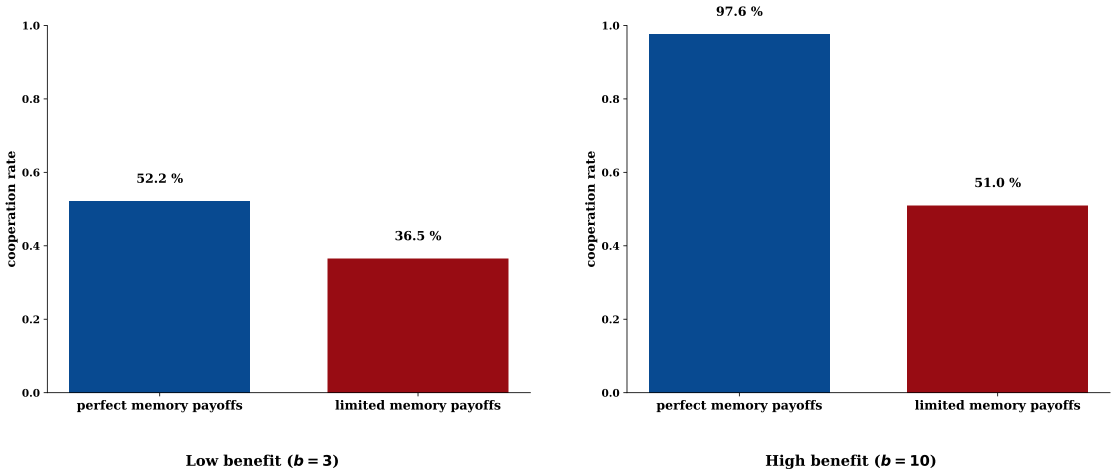


Figure 1: Evolutionary dynamics under expected payoffs and last round with one interaction payoffs. We have run two simulations of the evolutionary process described in section 2 for $t = 10^7$ time steps. For each time step, we have recorded the current resident population (y, p, q). Since simulations 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: $N = 100$, $c = 1$, $\beta = 1$, $\delta = 0.999$.



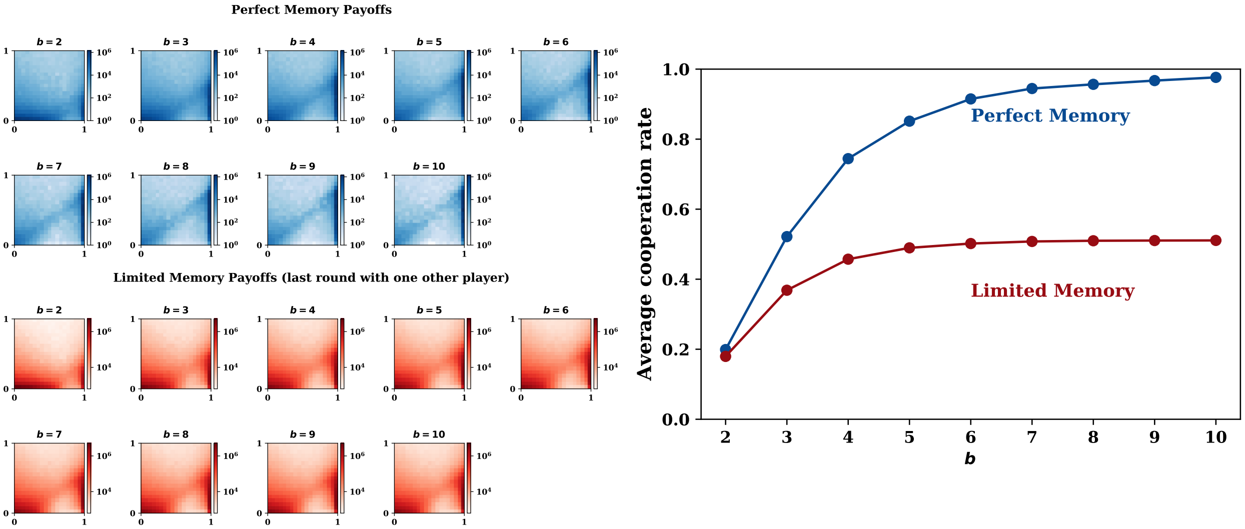


Figure 2: The evolution of cooperation for different benefit values. We vary the benefit of defection b. In all cases, expected payoffs appear to overestimate the average cooperation rate the population achieves. (**A**) the probabilities p, q for resident population over 10^7 time steps for each benefit 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 $t = 5 \times 10^7$ time steps for each parameter combination.

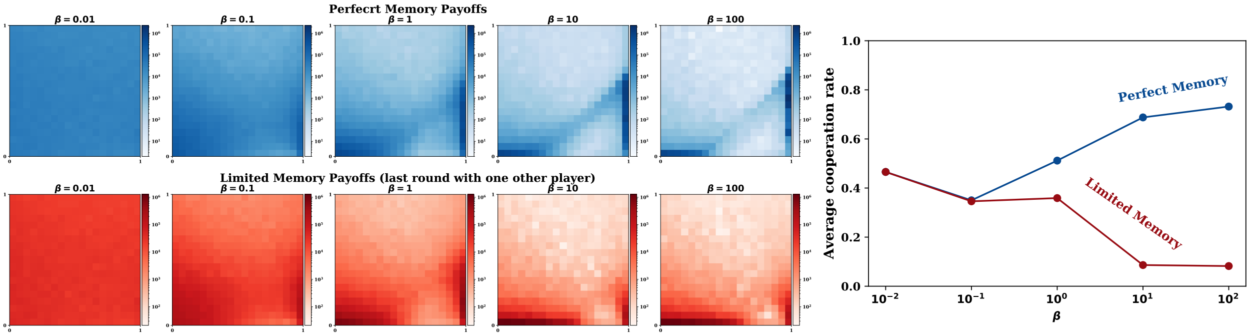


Figure 3: The evolution of cooperation for different selection strength values. We vary the selection strength β. In all cases, stochastic payoff evaluation tends to reduce the evolving cooperation rates. (**A**) the probabilities p, q for resident population over 10^7 time steps for each β value. (**B**) 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 $t = 5 \times 10^7$ time steps for each parameter combination.

4 Conclusions

Cooperation can be seen at odd, why is it that we choose to help others, increasing their payoff, at the expense of decreasing one's own? In spite of all the selfish genes' animal and human communities seem to altruistically help each other and cooperate, and evolutionary game theory has helped us shape our understanding of the evolution of cooperation.

Previous evolutionary models often feature a curious inconsistency. When modeling how individuals make decisions in each round, these models assume that players only remember the last round. However, when modeling how individuals update their strategies over time, individuals are assumed to have perfect memory.

Here, we have explored how robust cooperation is as models deviate from the perfect memory assumption. Initially we considered the donation game. We showed that when the last round payoff is used instead of the expected payoffs, cooperation can even evolve when individuals only use a minimum of information, however, the evolving cooperation rates are typically lower. The resident strategies were both less cooperative and less generous. This effect was only intensified we increase the benefit and the strength of selection independently. The results showed that as each parameter was being increased the difference between the cooperation rates were widening. This indicates that cooperative players benefit from being able to interact with everyone in the population.

We extended our approach and presented results not only based on different classes of social games, but also based on different limited memory payoffs. The analysis showed that specifically for the case of the prisoner's dilemma games the contrast between the two approaches can be significantly large. The prisoner's dilemma is one of the most applied types of social games when we discuss the evolution of cooperation. Our results indicate that cooperation struggles to evolve in the prisoner's dilemma when only a minimum of social information is used at the updating stage. Interestingly our simulations also showed that in same cases cooperation can benefit from limited memory. This was specific for the snowdrift game.

5 Acknowledgements

This work was supported by the European Research Council Starting Grant 850529: E-DIRECT.

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.

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. We present results for various values of $\mathcal{U}$ for all the symmetric 2×2 games.

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 strength 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 [7]. 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 $t <$ 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 [8]. 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,

$$M = \begin{bmatrix} p_1 p_2 & p_1 (1 - p_2) & p_2 (1 - p_1) & (1 - p_1)(1 - p_2) \\ p_2 q_1 & q_1 (1 - p_2) & p_2 (1 - q_1) & (1 - p_2)(1 - q_1) \\ p_1 q_2 & p_1 (1 - q_2) & q_2 (1 - p_1) & (1 - p_1)(1 - q_2) \\ q_1 q_2 & q_1 (1 - q_2) & q_2 (1 - q_1) & (1 - q_1)(1 - q_2) \end{bmatrix}. \tag{4}$$

whose stationary vector $\mathbf{v}$, combined with the payoff u, yields the game stage outcome for each strategy, $\langle \mathbf{v}(s_1, s_2), \mathbf{u} \rangle$ [9].

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 now explicitly defined 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 1.

Proposition 1. *Consider a repeated game, 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. Given that they can interact with only one other member of the population, they can interact either with each other or either can interact with another resident or with another mutant. Thus, in each updating stage there are five possible combinations of pairs (Figure 4).

Given the last round payoff and possible pair combinations for a single interaction, we define the probability that the respective last round payoffs of two players s_1, s_2 are given by u_1 and u_2 as,

$$\begin{aligned}
x(u_1,u_2) = \quad & \frac{1}{N-1} \cdot v_{u_1}(s_1,s_2) \cdot 1_{(u_1,u_2)\in\mathcal{U}_F^2} \\
& + \left(1 - \frac{1}{N-1}\right) \Big[\tfrac{k-1}{N-2}\tfrac{k-2}{N-3} v_{u_1}(s_1,s_2) v_{u_2}(s_2,s_2) + \tfrac{k-1}{N-2}\tfrac{N-k-1}{N-3} v_{u_1}(s_1,s_2) v_{u_2}(s_2,s_1) \\
& \qquad + \tfrac{N-k-1}{N-2}\tfrac{k-1}{N-3} v_{u_1}(s_1,s_1) v_{u_2}(s_2,s_2) + \tfrac{N-k-1}{N-2}\tfrac{N-k-2}{N-3} v_{u_1}(s_1,s_1) v_{u_2}(s_2,s_1) \Big].
\end{aligned} \tag{8}$$

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 $\frac{1}{(N-1)}$. 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

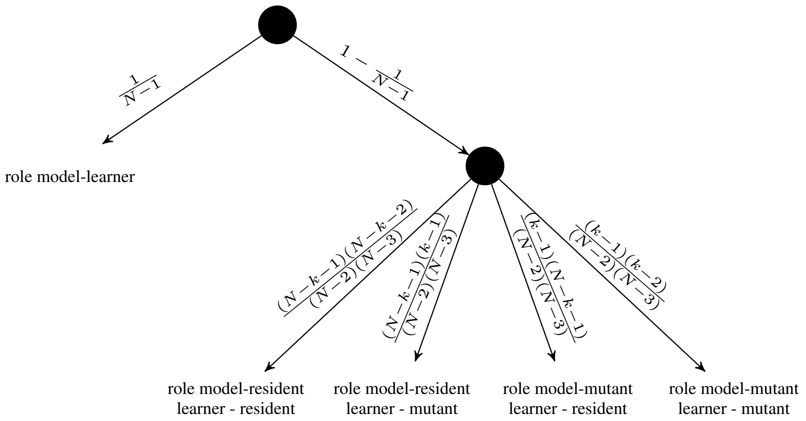

$\frac{1}{N-1}$
$1-\frac{1}{N-1}$
role model-learner
$\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(N-k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(k-2)}{(N-2)(N-3)}$
role model-resident learner - resident
role model-resident learner - mutant
role model-mutant learner - resident
role model-mutant learner - mutant

Figure 4: Possible pairings combination in the updating stage, given that individuals interact with only one other member in the population. At each step of the evolutionary process we choose a role model and a learner to update the population. We consider the case where both the role model and the learner estimate their fitness after interacting with a single member of the population. There are five possible pairings at each step. They interact with other with a probability $\frac{1}{N-1}$, and thus they do not interact with other with a probability $1-\frac{1}{N-1}$. In the latter case, each of them can interact with either a mutant or a resident. Both of them interact with a mutant with a probability $\frac{(k-1)(k-2)}{(N-2)(N-3)}$ and both interact with a resident with a probability $\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$. The last two possible pairings are that either of them interacts with a resident whilst the other interacts with a mutant, and this happens with a probability $\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$.

or a mutant.

Given that $N-k$ players use the resident strategy $s_R = (y_R, p_R, q_R)$ and that the remaining k players use the mutant strategy $s_M = (y_M, p_M, q_M)$, the probability that the number of mutants increases by one in one step of the evolutionary process can be written as

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

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

In this expression, $\frac{(N-k)}{N}$ is the probability that the randomly chosen learner is a resident, and $\frac{k}{N}$ 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_R and u_M that the two player may have received in their respective last rounds. We use $x(u_R, u_M)$ to denote the probability that the randomly chosen resident obtained a payoff of u_R in the last round of his respective game, and that the mutant obtained a payoff of u_M.

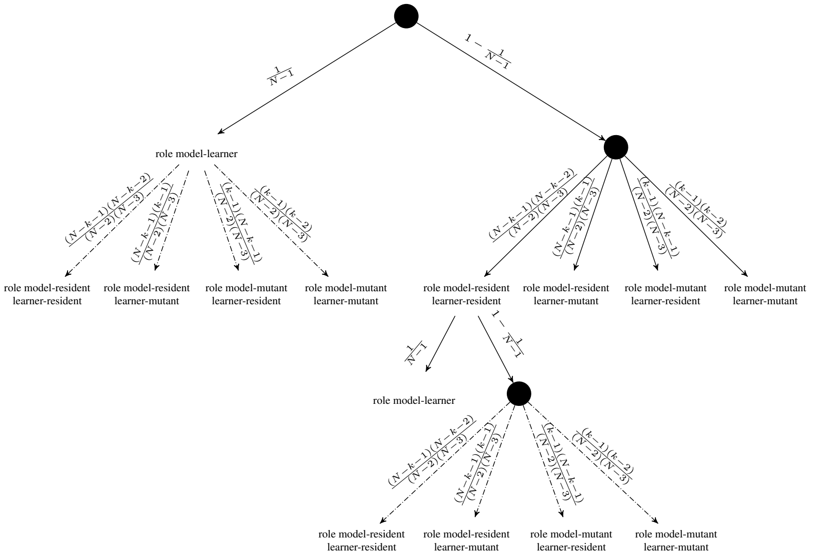
We extend our framework to consider the case where players update their strategies based on the outcome of **the last two turns and based on their interaction with two other members of the population**. At the stage game we define the payoff of a reactive strategy in the last two rounds, Proposition 2.

Proposition 2. *Consider a repeated game, 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. Let $\tilde{\mathcal{U}} = \{RR, RS, RT, RP, SR, SS, ST, SP, TR, TS, TT, TP, PR, PS, PT, PP\}$ denote the set of feasible payoffs in the last two rounds, and let $\tilde{\mathbf{u}}$ be the corresponding payoff vector. Then the probability that the s_1 player receives the payoff $u \in \tilde{\mathcal{U}}$ in the very last two rounds of the game is given by,*

$$\langle \tilde{\mathbf{v}}(s_1, s_2), \tilde{\mathbf{u}} \rangle, \ \textit{where } \tilde{\mathbf{v}} \in R^{16} \textit{ is given by },\tag{11}$$

$$\tilde{\mathbf{v}}(s_1, s_2) = (1-\delta) m_{a_1,a_2} \delta^2 \left[\mathbf{v_0}(I_4 - \delta M)^{-1}\right]_{a_1,a_2}, \quad m_{a_1,a_2} \in M \ \forall \ a_1, a_2 \in \{1,2,3,4\}.\tag{12}$$

In the updating stage we select a mutant and resident to be either the role model or the learner. Given that they can interact with two other members of the population there are a total of twenty four possible combinations of pairs (Figure 5).



$\frac{1}{N-1}$
$1 - \frac{1}{N-1}$
role model-learner
$\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(N-k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(k-2)}{(N-2)(N-3)}$
role model-resident learner-resident
role model-resident learner-mutant
role model-mutant learner-resident
role model-mutant learner-mutant
$\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(N-k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(k-2)}{(N-2)(N-3)}$
role model-resident learner-resident
role model-resident learner-mutant
role model-mutant learner-resident
role model-mutant learner-mutant
$\frac{1}{N-1}$
$1 - \frac{1}{N-1}$
role model-learner
$\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$
$\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(N-k-1)}{(N-2)(N-3)}$
$\frac{(k-1)(k-2)}{(N-2)(N-3)}$
role model-resident learner-resident
role model-resident learner-mutant
role model-mutant learner-resident
role model-mutant learner-mutant

Figure 5: Possible pairings combination in the updating stage, given that individuals interact with two other members in the population.

Given the last two rounds payoff and possible pair combinations with two members, we can define the probability x that the respective last round payoffs of two players s_1, s_2 are given by u_1 and u_2 similarly to Eq. (8). We follow the same approach to define the rest of the updating payoffs. These are the updating payoffs of the last round with two member, and the the last two rounds payoffs with one member.

Simulating the evolutionary process for more interactions and rounds quickly becomes computationally intractable. Our methodology could be extended to include n turns and m interactions. However, for the purpose of this work we explore the cases only up to two turns and two interactions.

B Evolutionary dynamics under limited memory payoffs for various social dilemmas

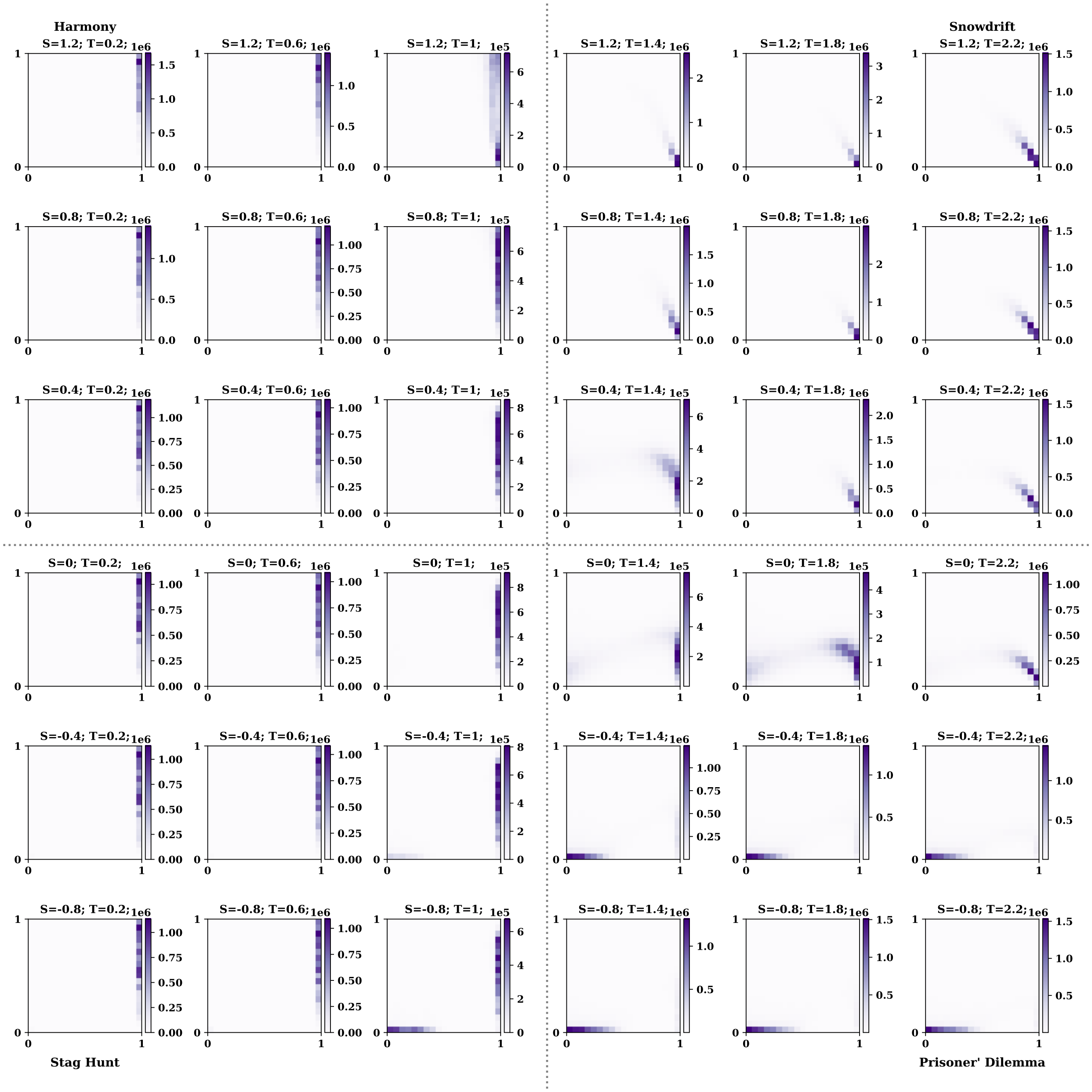


Figure 6: Evolutionary dynamics under last two rounds payoffs for various social dilemmas. We have run several simulations of the evolutionary process described in section 2 for $t = 10^7$ time steps. The graphs show how often the resident population chooses each combination (p, q) of conditional cooperation 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: $N = 100$, $\beta = 10$, $\delta = 0.999$.

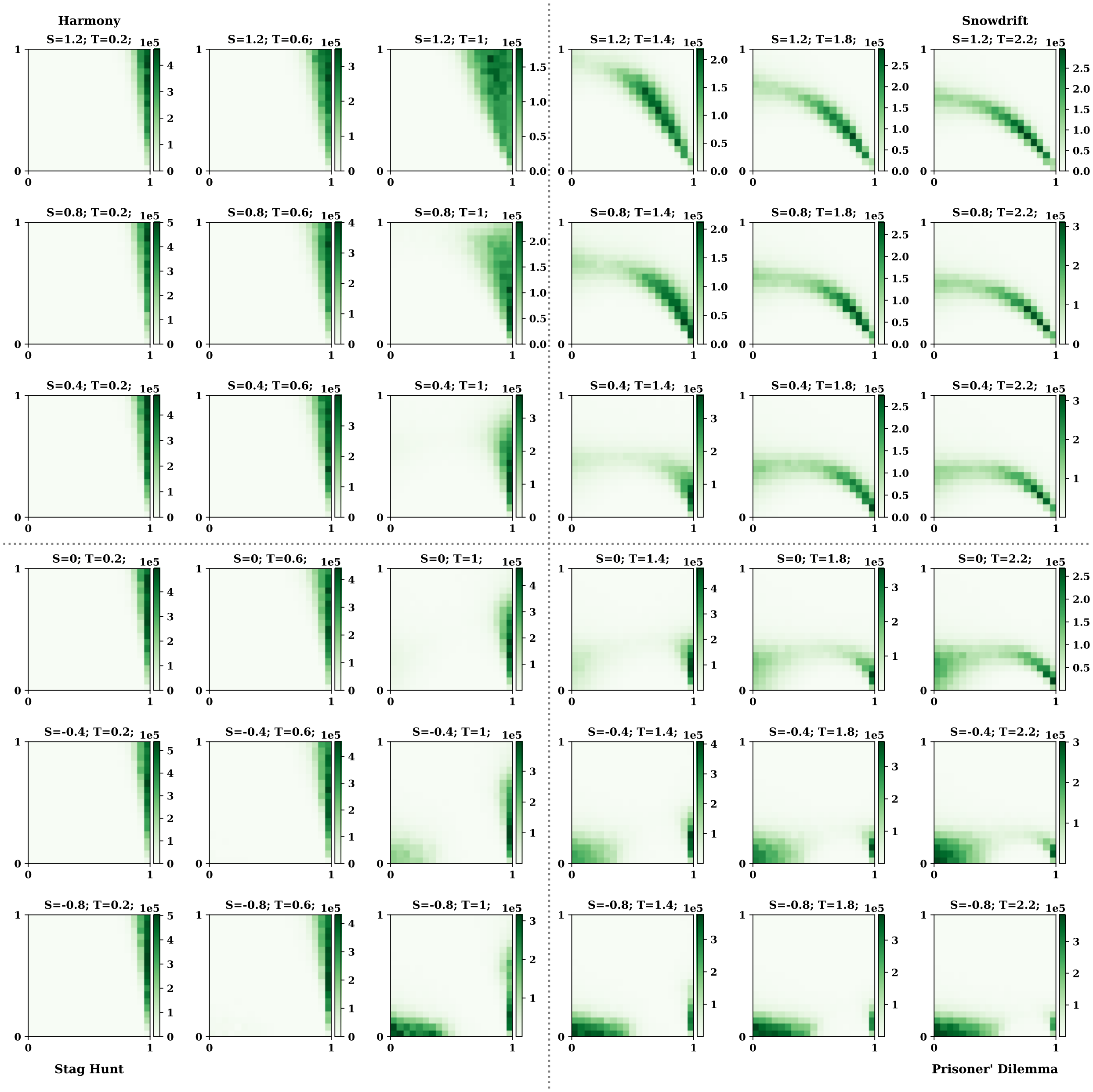


Figure 7: Evolutionary dynamics under last round payoffs with two members of the population for various social dilemmas. We have run several simulations of the evolutionary process described in section 2 for $t = 10^7$ time steps. The graphs show how often the resident population chooses each combination (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: $N = 100$, $\beta = 10$, $\delta = 0.999$.

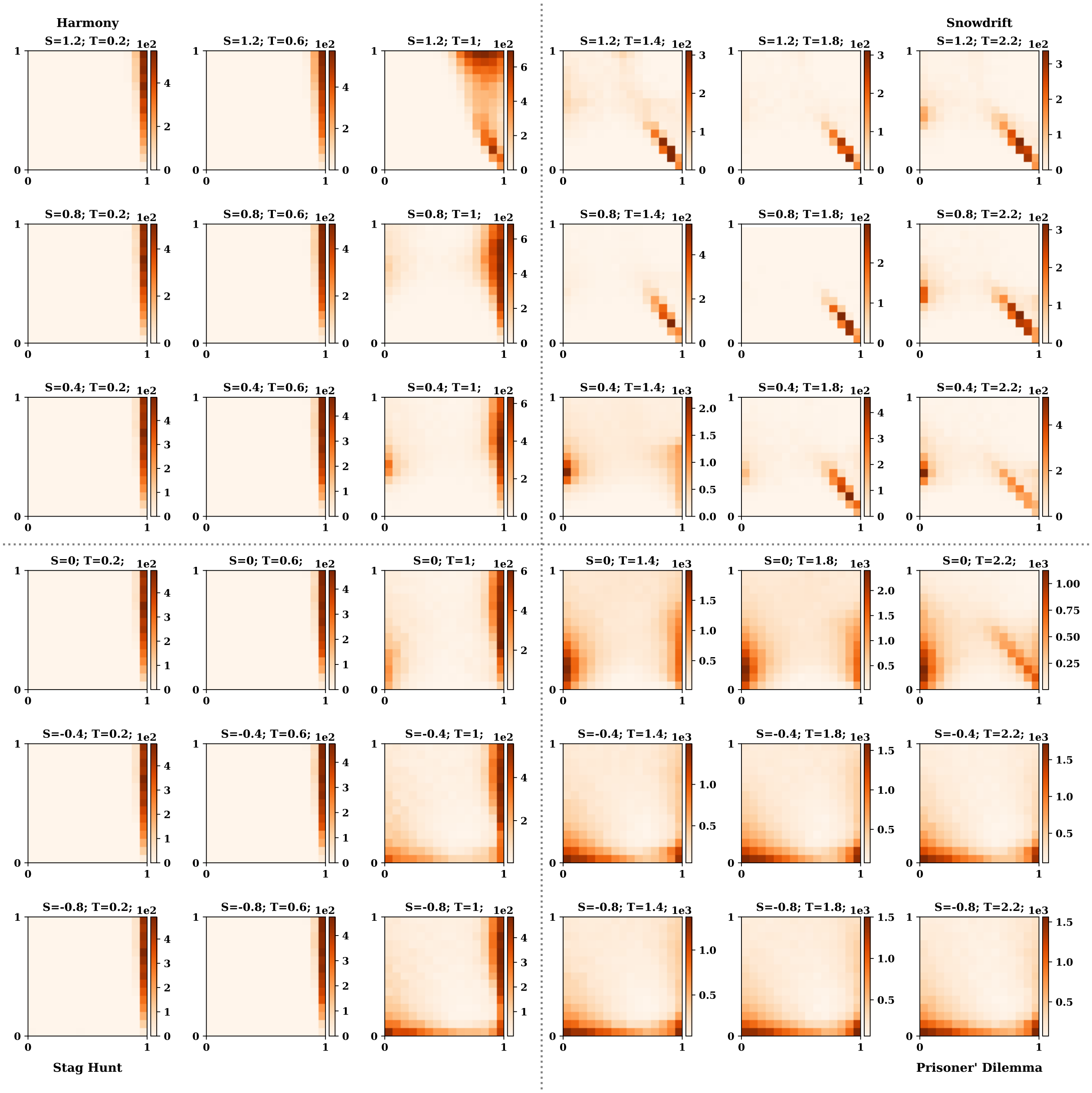


Figure 8: Evolutionary dynamics under last two rounds payoffs with two members of the population for various social dilemmas. We have run several simulations of the evolutionary process described in section 2 for $t = 10^7$ time steps. The graphs show how often the resident population chooses each combination (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: $N = 100$, $\beta = 10$, $\delta = 0.999$.

References

[1] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.

[2] Josef Hofbauer, Karl Sigmund, et al. *Evolutionary games and population dynamics*. Cambridge university press, 1998.

[3] Martin A Nowak and Karl Sigmund. Evolutionary dynamics of biological games. *science*, 303(5659):793–799, 2004.

[4] Christoph Hauert and György Szabó. Game theory and physics. *American Journal of Physics*, 73(5):405–414, 2005.

[5] Ross Cressman and Yi Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(Supplement 3):10810–10817, 2014.

[6] Arne Traulsen, Martin A Nowak, and Jorge M Pacheco. Stochastic dynamics of invasion and fixation. *Physical Review E*, 74(1):011909, 2006.

[7] 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.

[8] 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.

[9] 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.