

# Prisoner's dilemma with stochastic payoffs

## 1 Introduction

Evolutionary game theory consists of two main approaches. The first approach derives from the work of Maynard Smith and Price and employs the concept of an evolutionarily stable strategy. The second approach, constructs an explicit model of the process by which the frequency of strategies change in the population and studies properties of the evolutionary dynamics within that model.

The frequency of the strategies depends on a strategy's fitness, and commonly, the fitness of a strategy in a population is the expected payoff it receives from an interaction with a random member of the population.

The aim of this work is to understand the effect of the fitness function on the evolution of the population. More specifically, this work evaluates a number of different approaches for estimating the fitness of an individual.

## 2 Model

The Prisoner's Dilemma (PD) is a two player no-cooperative game. The players simultaneously and independently make a decision to Cooperate (C) or to defect (D). The payoffs for both players depend on their actions and the actions of their opponents. More specifically, the payoffs are given by:

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} \quad (1)$$

The payoffs,  $(R, P, S, T)$ , are constrained by  $T > R > P > S$  and  $2R > T + S$ .

A special case is the donation game," where each player can cooperate by providing a benefit  $b$  to the other player at their cost  $c$ , with  $0 < c < b$ . Then,  $T = b$ ,  $R = b - c$ ,  $S = -c$ ,  $P = 0$ , and matrix (1) is given by:

$$\begin{pmatrix} b - c & -c \\ b & 0 \end{pmatrix} \quad (2)$$

where  $c$  is the cost inflicted to an individual for cooperating and  $b$  is the benefit.

The questions of evolution become more interesting when repetition is considered, and the history can be accessed when making decisions. The iterated form of the game is called the Iterated Prisoner's Dilemma

(IPD). There are several two player repeated games such as the Snowdrift game [1], Harmony and the Stag Hunt [2] game. We present results in all four games.

The space of strategies in the IPD is infinite. To continue with our evolutionary analysis, we assume herein that individuals at most make use of simple **reactive strategies**. Reactive strategy are a set of memory-one strategies that only take into account the previous action of the opponent. An example of a reactive strategy is Tit For Tat. Reactive strategies 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.

A match between two reactive strategies takes the form of a stochastic process, and more specifically, of a Markov chain with four possible states  $CC, CD, DC, DD$  (the possible outcomes of each round). We assume a match between two reactive strategies  $s_1 = (y_1, p_1, q_1)$  and  $s_2 = (y_2, p_2, q_2)$ , the Markov process is described by the transition matrix  $M$ :

$$M = \begin{bmatrix} p_1 p_2 & p_1 (1 - p_2) & p_2 (1 - p_1) & (1 - p_1) (1 - p_2) \\ p_2 q_1 & q_1 (1 - p_2) & p_2 (1 - q_1) & (1 - p_2) (1 - q_1) \\ p_1 q_2 & p_1 (1 - q_2) & q_2 (1 - p_1) & (1 - p_1) (1 - q_2) \\ q_1 q_2 & q_1 (1 - q_2) & q_2 (1 - q_1) & (1 - q_1) (1 - q_2) \end{bmatrix} \quad (3)$$

The long run steady state probability vector  $\mathbf{v}$ , which is the solution to  $\mathbf{v}M = \mathbf{v}$ , can be combined with the payoff of matrix (1) (denoted as  $U$ ) to give the expected payoffs for each player. More specifically, the payoffs for a reactive strategy  $s_1$  against an opponent  $s_2$  is:

$$\mathbf{v}(s_1, s_2) \cdot U \quad (4)$$

where

$$U = \{R, S, T, P\}. \quad (5)$$

## 2.1 Evolutionary Dynamics

In evolution context we consider a population of  $N$  players, where  $N$  is even, and where mutations are sufficiently rare. At any point in time there are at most two different strategies are present in the population. Suppose there are  $N - k$  players who use the strategy  $s_1 = (y_1, p_1, q_1)$ , whereas  $k$  players use the strategy  $s_2 = (y_2, p_2, q_2)$ . We refer to these two player types as ‘residents’ and ‘mutants’, respectively.

Each step of the evolutionary process consists of two stages, a game stage and an updating stage.

1. In the game stage, each player is randomly matched with some other player in the population to interact

in one instance of the IPD.

2. In the updating stage, two players are randomly drawn from the population, a ‘learner’ and a ‘exemplar’. Given that the learner’s payoff in the last round is  $u_L \in \mathcal{U}$  and that the exemplar’s last round’s payoff  $u_E \in \mathcal{U}$ , we assume the learner adopts the role model’s strategy with probability

$$\rho(u_L, u_E) = \frac{1}{1 + \exp[-\beta(u_E - u_L)]}. \quad (6)$$

where  $\beta \geq 0$  corresponds to the strength of selection.

We iterate this basic evolutionary step until either the mutant strategy goes extinct, or until it fixes in the population (in which case the mutant strategy becomes the new resident strategy). After either outcome, we introduce a new mutant strategy  $s'_2 = (y'_2, p'_2, q'_2)$  (uniformly chosen from all reactive strategies at random), and we set the number of mutants to  $k = 1$ . This process of mutation and fixation/extinction is then iterated many times.

We compare this process for what we defined as **stochastic payoff evaluation** with the analogous process where players update their strategies with respect to their **expected** payoffs,

$$\begin{aligned} \pi_1 &= \frac{N-k-1}{N-1} \cdot \langle \mathbf{v}(s_1, s_1), \mathbf{U} \rangle + \frac{k}{N-1} \cdot \langle \mathbf{v}(s_1, s_2), \mathbf{U} \rangle, \\ \pi_2 &= \frac{N-k}{N-1} \cdot \langle \mathbf{v}(s_2, s_1), \mathbf{U} \rangle + \frac{k-1}{N-1} \cdot \langle \mathbf{v}(s_2, s_2), \mathbf{U} \rangle. \end{aligned} \quad (7)$$

In the limit of no discounting,  $\delta \rightarrow 1$ , this process based on expected payoffs has been considered in [3].

## 2.2 Stochastic payoff evaluation

We define **stochastic payoff** as the average payoff  $u \in \mathcal{U}$  a player receives in the last  $n$  rounds of the game given that they interact with  $m$  players.

**Case  $n = m = 1$ .**

Initially, consider the situation where  $n = m = 1$ . The player’s stochastic payoff is what they receive in the last round against a single opponent. There only four possible outcomes for the last round, those are  $CC, CD, DC, DD$ . Consider two players with reactive strategies  $S_1 = (y_1, p_1, q_1)$  and  $S_2 = (y_2, p_2, q_2)$  who interact in a repeated prisoner’s dilemma with continuation probability  $\delta$ , the probability that are in each of the four possible states in the last round is given by:

$$\mathbf{v}(s_1, s_2) = \left( \mathbf{v}_R(s_1, s_2), \mathbf{v}_S(s_1, s_2), \mathbf{v}_T(s_1, s_2), \mathbf{v}_P(s_1, s_2) \right). \quad (8)$$

where,

$$\begin{aligned}
\mathbf{v}_R(S_1, S_2) &= (1-\delta) \frac{y_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( q_1 + r_1((1-\delta)y_2 + \delta q_2) \right) \left( q_2 + r_2((1-\delta)y_1 + \delta q_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\
\mathbf{v}_S(S_1, S_2) &= (1-\delta) \frac{y_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( q_1 + r_1((1-\delta)y_2 + \delta q_2) \right) \left( \bar{q}_2 + \bar{r}_2((1-\delta)y_1 + \delta p_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\
\mathbf{v}_T(S_1, S_2) &= (1-\delta) \frac{\bar{y}_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( \bar{q}_1 + \bar{r}_1((1-\delta)y_2 + \delta p_2) \right) \left( q_2 + r_2((1-\delta)y_1 + \delta q_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\
\mathbf{v}_P(S_1, S_2) &= (1-\delta) \frac{\bar{y}_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( \bar{q}_1 + \bar{r}_1((1-\delta)y_2 + \delta p_2) \right) \left( \bar{q}_2 + \bar{r}_2((1-\delta)y_1 + \delta p_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}.
\end{aligned} \tag{9}$$

*Proof.* Assume a repeated prisoner's dilemma between two reactive strategies. Given the continuation probability  $\delta$ , probability that the game ends in the after the first round  $(1 - \delta)$  and the expected distribution of the four outcomes in the very first round is  $\mathbf{v}_0$  defined as. Following the first round the, the outcome of the next rounds with a probability  $\delta$  is  $M$  such that,

...

It can shown that,  $!(1-\delta)\mathbf{v}_0(I_4 - \delta M)^{-1}$  and with some algebraic manipulation we derive to Equation 11.  $\square$

is given by  $v_u(S_1, S_2)$

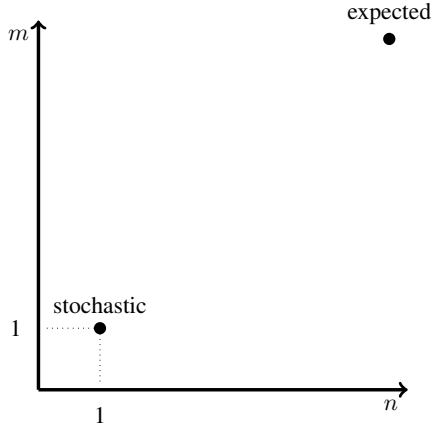
Consider two players with reactive strategies  $S_1 = (y_1, p_1, q_1)$  and  $S_2 = (y_2, p_2, q_2)$  who interact in a repeated prisoner's dilemma with continuation probability  $\delta$ . We consider the vector

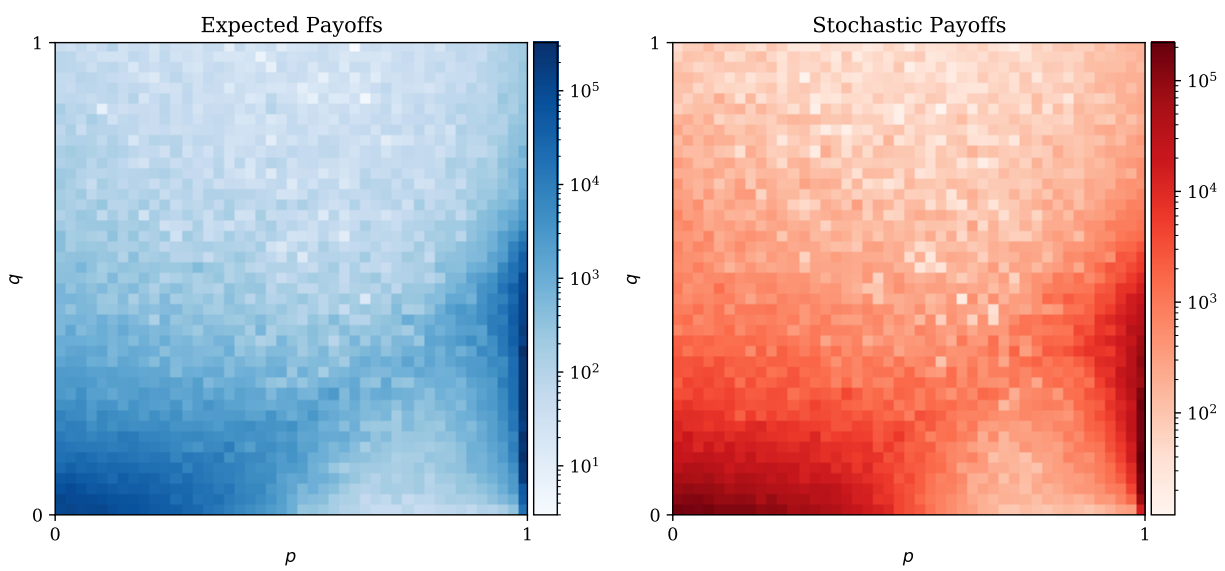
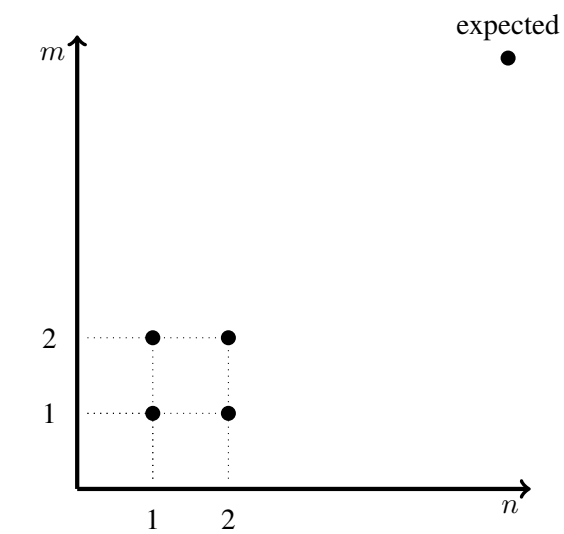
$$\mathbf{v}(S_1, S_2) = \left( v_R(S_1, S_2), v_S(S_1, S_2), v_T(S_1, S_2), v_P(S_1, S_2) \right) := (1-\delta)\mathbf{v}_0(I_4 - \delta M)^{-1}. \tag{10}$$

Here,  $\mathbf{v}_0$  denotes the expected distribution of the four outcomes in the very first round,  $I_4$  is the  $4 \times 4$  identity matrix, and  $M$  is the transition matrix of the Markov chain. The entries of  $\mathbf{v}$  can be calculated explicitly,

$$\begin{aligned}
v_R(S_1, S_2) &= (1-\delta) \frac{y_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( q_1 + r_1 ((1-\delta)y_2 + \delta q_2) \right) \left( q_2 + r_2 ((1-\delta)y_1 + \delta q_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\
v_S(S_1, S_2) &= (1-\delta) \frac{y_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( q_1 + r_1 ((1-\delta)y_2 + \delta q_2) \right) \left( \bar{q}_2 + \bar{r}_2 ((1-\delta)y_1 + \delta p_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\
v_T(S_1, S_2) &= (1-\delta) \frac{\bar{y}_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( \bar{q}_1 + \bar{r}_1 ((1-\delta)y_2 + \delta p_2) \right) \left( q_2 + r_2 ((1-\delta)y_1 + \delta q_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\
v_P(S_1, S_2) &= (1-\delta) \frac{\bar{y}_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left( \bar{q}_1 + \bar{r}_1 ((1-\delta)y_2 + \delta p_2) \right) \left( \bar{q}_2 + \bar{r}_2 ((1-\delta)y_1 + \delta p_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}.
\end{aligned} \tag{11}$$

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  $i \in \{1, 2\}$ . 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. Then one can show the following result.





**Figure 1:** Blah blah.

### 3 Pairwise imitation dynamics under stochastic payoff evaluation

#### 3.1 Basic setup

In the following we consider a population of size  $N$ , where  $N$  is even. We assume mutations are sufficiently rare such that at any point in time at most two different strategies are present in the population. Suppose there are  $N - k$  players who use the strategy  $S_1 = (y_1, p_1, q_1)$ , whereas  $k$  players use the strategy  $S_2 = (y_2, p_2, q_2)$ . We refer to these two player types as ‘residents’ and ‘mutants’, respectively.

Each step of the evolutionary process consists of two stages, a game stage and an updating stage. In the game stage, each player is randomly matched with some other player in the population to interact in one instance of the repeated prisoner’s dilemma. In the updating stage, two players are randomly drawn from the population, a ‘learner’ and a ‘role model’. Given that the learner’s payoff in the last round is  $u_L \in \mathcal{U}$  and that the role model’s last round’s payoff  $u_M \in \mathcal{U}$ , we assume the learner adopts the role model’s strategy with probability

$$\rho(u_L, u_M) = \frac{1}{1 + \exp[-\beta(u_M - u_L)]}. \quad (12)$$

The parameter  $\beta \geq 0$  corresponds to the strength of selection.

We iterate this basic evolutionary step until either the mutant strategy goes extinct, or until it fixes in the population (in which case the mutant strategy becomes the new resident strategy). After either outcome, we introduce a new mutant strategy  $S'_2 = (y'_2, p'_2, q'_2)$  (uniformly chosen from all reactive strategies at random), and we set the number of mutants to  $k = 1$ . This process of mutation and fixation/extinction is then iterated many times.

We compare this process for stochastic payoff evaluation with the analogous process where players update their strategies with respect to their *expected* payoffs,

$$\begin{aligned} \pi_1 &= \frac{N-k-1}{N-1} \cdot \langle \mathbf{v}(S_1, S_1), \mathbf{u} \rangle + \frac{k}{N-1} \cdot \langle \mathbf{v}(S_1, S_2), \mathbf{u} \rangle, \\ \pi_2 &= \frac{N-k}{N-1} \cdot \langle \mathbf{v}(S_2, S_1), \mathbf{u} \rangle + \frac{k-1}{N-1} \cdot \langle \mathbf{v}(S_2, S_2), \mathbf{u} \rangle. \end{aligned} \quad (13)$$

In the limit of no discounting,  $\delta \rightarrow 1$ , this process based on expected payoffs has been considered in [3].

#### 3.2 Fixation probabilities under stochastic payoff evaluation

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

$$\lambda_k^+ = \frac{N-k}{N} \cdot \frac{k}{N} \cdot \sum_{u_1, u_2 \in \mathcal{U}} x(u_1, u_2) \cdot \rho(u_1, u_2). \quad (14)$$

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  $u_2$ . Given that the payoffs are  $u_1$  and  $u_2$ , the imitation probability is then given by  $\rho(u_1, u_2)$ , as specified by Eq. (12). The probability that the respective payoffs of the players are given by  $u_1$  and  $u_2$  can be calculated as

$$\begin{aligned} x(u_1, u_2) = & \frac{1}{N-1} \cdot v_{u_1}(S_1, S_2) \cdot 1_{(u_1, u_2) \in \mathcal{U}_F^2} \\ & + \left(1 - \frac{1}{N-1}\right) \left[ \frac{k-1}{N-2} \frac{k-2}{N-3} v_{u_1}(S_1, S_2) v_{u_2}(S_2, S_2) + \frac{k-1}{N-2} \frac{N-k-1}{N-3} v_{u_1}(S_1, S_2) v_{u_2}(S_2, S_1) \right. \\ & \left. + \frac{N-k-1}{N-2} \frac{k-1}{N-3} v_{u_1}(S_1, S_1) v_{u_2}(S_2, S_2) + \frac{N-k-1}{N-2} \frac{N-k-2}{N-3} v_{u_1}(S_1, S_1) v_{u_2}(S_2, S_1) \right]. \end{aligned} \quad (15)$$

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,  $(u_1, u_2) \in \mathcal{U}_F^2 := \{(R, R), (S, T), (T, S), (P, P)\}$ , 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.

Analogously, we can calculate the probability that the number of mutants decreases by one in one step of the evolutionary process. This probability is

$$\lambda_k^- = \frac{N-k}{N} \cdot \frac{k}{N} \cdot \sum_{u_1, u_2 \in \mathcal{U}} x(u_1, u_2) \cdot \rho(u_2, u_1). \quad (16)$$

The fixation probability of the mutant strategy then takes the standard form [4],

$$\varphi = \frac{1}{1 + \sum_{i=1}^{N-1} \prod_k^i \frac{\lambda_k^-}{\lambda_k^+}}. \quad (17)$$



### 3.3 Invasion analysis of ALLD into GTFT

In the following, we apply the above formalism to calculate how easily a single ALLD mutant can invade into a resident population with strategy GTFT. In that case,  $S_1 = (1, 1, q)$ ,  $S_2 = (0, 0, 0)$ , and  $k = 1$ . When two GTFT players interact in the game, their respective probabilities for each of the four outcomes in the last round simplify to

$$\begin{aligned} v_R(GTFT, GTFT) &= 1, & v_S(GTFT, GTFT) &= 0, \\ v_T(GTFT, GTFT) &= 0, & v_P(GTFT, GTFT) &= 0. \end{aligned} \quad (18)$$

On the other hand, if an ALLD player interacts with a GTFT player, the respective probabilities according to Eq. (11) become

$$\begin{aligned} v_R(ALLD, GTFT) &= 0, & v_S(ALLD, GTFT) &= 0, \\ v_T(ALLD, GTFT) &= 1 - \delta + \delta q, & v_P(ALLD, GTFT) &= \delta(1 - q). \end{aligned} \quad (19)$$

As a consequence, we obtain the following probabilities  $x(u_1, u_2)$  that the payoff of a randomly chosen GTFT player is  $u_1$  and that the payoff of the ALLD player is  $u_2$ ,

$$\begin{aligned} x(R, T) &= \frac{N-2}{N-1} \cdot (1 - \delta + \delta q), \\ x(R, P) &= \frac{N-2}{N-1} \cdot \delta(1 - q), \\ x(S, T) &= \frac{1}{N-1} \cdot (1 - \delta + \delta q), \\ x(P, P) &= \frac{1}{N-1} \cdot \delta(1 - q). \\ x(u_1, u_2) &= 0 \quad \text{for all other payoff pairs } (u_1, u_2). \end{aligned} \quad (20)$$

As a consequence, we can calculate the ratio of transition probabilities as

$$\frac{\lambda_1^+}{\lambda_1^-} = \frac{\frac{N-2}{N-1} \cdot \left( \frac{1 - \delta + \delta q}{1 + \exp[-\beta(T-R)]} + \frac{\delta(1-q)}{1 + \exp[-\beta(P-R)]} \right) + \frac{1}{N-1} \left( \frac{1 - \delta + \delta q}{1 + \exp[-\beta(T-S)]} + \frac{\delta(1-q)}{2} \right)}{\frac{N-2}{N-1} \cdot \left( \frac{1 - \delta + \delta q}{1 + \exp[-\beta(R-T)]} + \frac{\delta(1-q)}{1 + \exp[-\beta(R-P)]} \right) + \frac{1}{N-1} \left( \frac{1 - \delta + \delta q}{1 + \exp[-\beta(S-T)]} + \frac{\delta(1-q)}{2} \right)}. \quad (21)$$

In particular, in the limit of strong selection  $\beta \rightarrow \infty$  and large populations  $N \rightarrow \infty$ , we obtain

$$\frac{\lambda_1^+}{\lambda_1^-} = \frac{1 - \delta + \delta q}{\delta(1 - q)}. \quad (22)$$

This ratio is smaller than 1 (such that ALLD is disfavored to invade) if  $q < 1 - 1/(2\delta)$ . For infinitely repeated

games,  $\delta \rightarrow 1$ , this condition becomes  $q < 1/2$  (for  $q = 1/2$ , the payoff of the ALLD player is  $T > R$  for half of the time, and it is  $P < R$  for the other half. The probability that the number of mutants increase by one equals the probability that the mutant goes extinct).

## References

- [1] Robert Sugden et al. *The economics of rights, co-operation and welfare*. Springer, 2004.
- [2] Brian Skyrms. The stag hunt. In *Proceedings and Addresses of the American Philosophical Association*, volume 75, pages 31–41. JSTOR, 2001.
- [3] 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.
- [4] 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.