Evolutionary stability of first-order-information indirect reciprocity in sizable groups

Shinsuke Suzuki [a,*] , Eizo Akiyama [b]

[a] *Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-0006, Japan*

[b] *Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-0006, Japan*

Abstract

Indirect reciprocity is considered as a key mechanism for explaining the evolution of cooperation in situations where the same individuals interact only a few times. Under indirect reciprocity, an individual who helps others gets returns *indirectly* from others who know her good *reputation*. Recently, many studies have discussed the effect of reputation criteria based only on the former actions of the others (first-order information) and of those based also on the former actions of opponents of the others (second-order information) on the evolution of indirect reciprocity. In this study, we investigate the evolutionary stability of the indirectly reciprocal strategy (discriminating strategy: DIS), which cooperates only with opponents who have good reputations, in the $n(> 2)$-person case where more than two persons take part in a single *group* (interaction). We show that in the n-person case, DIS is an evolutionarily stable strategy (ESS) even under the image-scoring reputation criterion, which is based only on first-order information and where cooperations (defections) are judged to be good (bad). This result is in contrast to that of the two-person case where DIS is not an ESS under reputation criteria based on first-order information and where evolutionary stability of DIS requires both first- and second-order information.

Key words: ESS, indirect reciprocity, image scoring, reputation, prisoner's dilemma

* Corresponding author. Tel: +81-298-53-5571.
Email address: suzuki92@sk.tsukuba.ac.jp (Shinsuke Suzuki).

1 Introduction

Indirect reciprocity is considered to be a key mechanism for explaining the evolution of cooperation in situations where the same individuals interact only a few times and where the reputations of individuals affects the decision-making process (Alexander, 1987; Nowak & Sigmund, 2005). Under indirect reciprocity, an individual who helps others gets returns *indirectly* from others who know her good reputation in the community.

Nowak & Sigmund (1998a,b) formalized a mathematical model of indirect reciprocity as an evolutionary *two*-person giving game involving *image scoring* that evaluates the reputations of individuals in a community. In their model, pairs of individuals interact only a few times and all individuals are informed about their partners' reputations (*image score*), which reflects the partners' past behaviors. They showed that, in this model, an indirectly reciprocal strategy, called *discriminating strategy* (DIS), which cooperates only with opponents who have good reputations, can prevail in the population. Since the seminal works by Nowak & Sigmund, many theoretical and experimental studies on indirect reciprocity have been conducted (e.g. Leimar & Hammerstein, 2001; Wedekind & Milinski, 2000; Lotem et al., 1999; Milinski et al., 2001, 2002b; Panchanathan & Boyd, 2003, 2004; Fishman, 2003; Mohtashemi & Mui, 2003; Ohtsuki & Iwasa, 2004; Brandt & Sigmund, 2004, 2005; Takahashi & Mashima, 2006).

One of the most important issues regarding the studies of indirect reciprocity has been how people judge (define) others' goodness (e.g. Leimar & Hammerstein, 2001; Milinski et al., 2001; Panchanathan & Boyd, 2003; Ohtsuki & Iwasa, 2004, 2005; Brandt & Sigmund, 2004; Bolton et al., 2005; Takahashi & Mashima, 2006). Put more precisely, a controversial issue in indirect reciprocity is whether people use only first-order information, which are the former actions of the others, or use in addition second-order information, which are the former actions of opponents of the others. Using second-order information, each individual can distinguish between defections to former defectors and those to former cooperators. In other words, second-order information enables us to distinguish between justified and unjustified defections. (The image scoring given in Nowak & Sigmund (1998a,b) is a reputation criterion based on first-order information, i.e., others' past behavior.)

Many theoretical studies (e.g. Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2003; Ohtsuki & Iwasa, 2004, 2005; Brandt & Sigmund, 2004), using the *two*-person game, have pointed out that, under image scoring that is based on first-order information, the indirectly reciprocal strategy, DIS, is not an *evolutionarily stable strategy (ESS)*. The reason that DIS is not an ESS is as follows: DIS strategists hurt each other in response to error de-

fections because a DIS strategist does not cooperate with individuals who defected out of error, causing her own reputation to become bad. As a result, DIS strategists defect against each other. On the contrary, individuals who adopt an unconditionally cooperative strategy called $ALLC$ always intend to cooperate. Therefore $ALLC$ strategists do not hurt others and are not hurt by DIS strategists. Consequently, in a population consisting of DIS, the fitness of DIS is less than $ALLC$, so the DIS population can be invaded by $ALLC$. On the other hand, several studies (e.g. Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2003; Brandt & Sigmund, 2004; Takahashi & Mashima, 2006) have stated that DIS is an ESS under a reputation criterion that reflects not only first-order information but also second-order information. Especially, Panchanathan & Boyd (2003) showed that, in the presence of implementation error, DIS is an ESS under *standing* reputation criterion under which an individual's defections to former defectors are justified (Sugden, 1986). Moreover, Ohtsuki & Iwasa (2004, 2005) showed that, in the presence of implementation error and objective perception error, under the eight reputation criteria called the *leading eight*, which include standing, under which an individual's defections to former defectors is justified, a strategy forming a cooperative society is an ESS. Furthermore, Takahashi & Mashima (2006) showed that, in the presence of subjective perception error, DIS is an ESS under the reputation criteria under which an individual's cooperations with former defectors is unjustified [1].

Note that most of the above studies on indirect reciprocity have presumed dyadic interaction. However, in the real world, more than two persons often take part in a single *group* (interaction). The evolutionary stability of DIS, which cooperates only when all the opponents in the group have good reputations, was investigated for groups of various sizes in Suzuki & Akiyama (submitted) [2]. They showed that in a population with DIS, $ALLC$, and $ALLD$ (unconditionally defective strategy), DIS is an ESS not only under standing, which reflects second-order information, but also under image scoring, which reflects only first-order information, if the group size is more than two and if the cost-to-benefit ratio of cooperation is sufficiently small.

In this study, we analyze the evolutionary stability of DIS under image scoring in the situation where more than two persons take part in a single group and where the population consists of all the possible strategies that decide their own action based on the number of opponents in the group whose reputation

[1] However, experimental studies have not been able to conclude whether people actually use only first-order information or also use second-order information. For instance, Milinski et al. (2001) found that people do not actually use second-order information, while Bolton et al. (2005) found that they do.

[2] Furthermore Suzuki & Akiyama (2005) used agent-based computer simulation to investigate the evolution of indirect reciprocity in groups of various sizes. However they did not mention the evolutionary stability.

is good. Clearly, in the absence of error, DIS is not an ESS because $ALLC$ is an alternative best reply to DIS. Therefore, in this study, we focus on the case with implementation error. First, we investigate the evolutionary stability of DIS analytically, and next, compare the result with that derived by numerical calculation.

2 Indirect reciprocity in groups of various sizes

Consider a population composed of an infinite number of individuals. Each individual in the population has her own *reputation* either *G (good)* or *B (bad)*.

Each *generation* comprises a number of *rounds*. After the first round, each of the subsequent rounds occurs with probability w ($0 \ll w < 1$), i.e., the expected value of the number of rounds in a generation is $1/(1-w)$. At the beginning of each generation, the reputation of each individual is G.

In each round, all individuals are divided randomly into groups, each of which comprises $n(\geq 3)$ individuals, and play an n-person prisoner's dilemma game in each group. In this game, each of the individuals chooses either to "cooperate (C)" or "defect (D)". The payoffs for a cooperator, $V(\,C\,|\,k)$, and that for a defector, $V(\,D\,|\,k)$, where k is the number of opponents cooperating in the group, satisfy the following conditions (e.g. Boyd & Richerson, 1988; Molander, 1992; Eriksson & Lindgren, 2005):

- Each individual is better off choosing D regardless of the choices of the other members in the group (dominance of D):

$$V(D\,|\,k) > V(C\,|\,k). \tag{1}$$

- The payoff for each individual increases as the number of cooperators increases (monotonicity):

$$V(D\,|\,k) > V(D\,|\,k-1),\ V(C\,|\,k) > V(C\,|\,k-1). \tag{2}$$

- The average payoff of individuals in the group increases with the number of cooperators (efficiency of cooperation):

$$\begin{aligned}&(k+1)V(C\,|\,k) + (n-k-1)V(D\,|\,k+1) > \\ &kV(C\,|\,k-1) + (n-k)V(D\,|\,k).\end{aligned} \tag{3}$$

In this paper, we assume that the payoff functions for cooperators and defectors are, as in Joshi (1987); Eriksson & Lindgren (2005); Lindgren & Johansson (2001), etc., calculated as a linear combination of the payoffs against the $n-1$

opponents in *two*-person prisoner's dilemma games whose payoff is given in Table 1:

$$V(C \mid k) = \frac{bk}{n-1} - c \tag{4}$$

$$V(D \mid k) = \frac{bk}{n-1}, \tag{5}$$

where $b > c > 0$ and we have divided the leaner combination of the payoffs by $n-1$ in order to compare results from different group sizes. Of course, these payoff functions satisfy the conditions (1)-(3). In this game, b represents the benefit of cooperation to the recipients and c denotes the cost to the donor. Note that this game is a straightforward expansion of the two-person giving game used in many studies regarding indirect reciprocity (e.g. Nowak & Sigmund, 1998a,b; Panchanathan & Boyd, 2003). In the case of $n = 2$, the payoff functions (4)-(5) are the same as those in Nowak & Sigmund (1998a,b); Panchanathan & Boyd (2003).

Table 1
Payoff of *two*-person prisoner's dilemma game ($b > c > 0$).

		Player 2	
		cooperation	defection
Player 1	cooperation	$(b-c,\ b-c)$	$(-c,\ b)$
	defection	$(b,\ -c)$	$(0,\ 0)$

Moreover, implementation error is introduced by the parameter ϵ ($0 < \epsilon \ll 1$). With a probability ϵ, an individual who intends to cooperate fails to cooperate due to a lack of resources or a mistake [3]. In other words, an individual who intends to cooperate succeeds in cooperation with a probability $\hat{\epsilon} = 1 - \epsilon$ ($0 \ll \hat{\epsilon} < 1$). In this study, we mainly use the probability of the success, $\hat{\epsilon}$.

2.1 Reputation criterion

In this model, the reputation of opponents affects the decision-making process. How is the reputation assigned to each individual? In this study, we adopt "*image scoring*" as a *reputation criterion*, which prescribes how to judge the

[3] We, as in Panchanathan & Boyd (2003); Fishman (2003); Brandt & Sigmund (2004), do not consider errors that cause unintentional cooperation, i.e., an individual who intends to defect never fails to defect. Furthermore, perception error (Ohtsuki & Iwasa, 2004; Takahashi & Mashima, 2006) is not included.

reputation of others on the basis of the others' past action. Under the criterion, which was first used in Nowak & Sigmund (1998a,b), the reputation of an individual who has defected becomes B and that of an individual who has cooperated becomes G. This is a simple reputation criterion that requires only the past action of opponents (first-order information) (cf., standing reputation criterion given in Leimar & Hammerstein (2001); Panchanathan & Boyd (2003) uses second-order information).

2.2 Strategies

In this study, individuals are supposed to use a strategy that prescribes how to decide their own action based on their opponents' reputations.

Such a strategy is represented by a vector $\mathbf{p} = (p_0, \cdots, p_{n-1})$, where $p_i \in \{0, 1\}$, in which p_i represents an action of the individual when the number of opponents who have reputation G is i, and 0 and 1 mean defection and cooperation, respectively. For example, in the four-person case, a $\mathbf{p} = (0, 1, 0, 0)$ strategist cooperates only when one opponent has reputation G, and a $\mathbf{p} = (0, 0, 1, 1)$ strategist cooperates only when more than one opponents have reputation G. In other words, there are 2^n different strategies.

3 Evolutionary stability of indirect reciprocity (analytical investigation)

Here we discuss the evolutionary stability of $\hat{\mathbf{p}} = (0, ..., 0, 1)$, which is called discriminating strategy (DIS).

For this purpose, we consider the situation where a small number of $\mathbf{p}$ strategist enter a population dominated by $\hat{\mathbf{p}}$ strategists and where the frequency of the invader is sufficiently small for us to presume that each individual always interacts only with incumbent $\hat{\mathbf{p}}$ strategists. In the rest of this paper, we always consider this situation.

Let $W(\mathbf{p}|\hat{\mathbf{p}})$ be the fitness in this situation. In this section, we demonstrate the inequality

$$W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}|\hat{\mathbf{p}}) \quad (for\ all\ \mathbf{p} \neq \hat{\mathbf{p}}). \tag{6}$$

If the inequality holds, $\hat{\mathbf{p}}$ is an ESS.

3.1 Frequency of individuals with reputation G

With $\hat{g}_t$, we represent the frequency of individuals with reputation G among incumbent $\hat{\mathbf{p}}$ strategists, and with g_t, that among invader $\mathbf{p}$ strategists. Under image scoring, only the individuals who have cooperated at round $t-1$ have reputation G at round t. In other words, to derive the frequency of individuals with reputation G, it is necessary to derive the probability that each individual cooperates.

The possible situations that each individual faces are the following n cases: there is no opponent with G reputation; there is one; $\cdots$; and there are $n-1$ opponents with G reputation. Since we are considering the case where each individual always interacts with only incumbent $\hat{\mathbf{p}}$ strategists, the probability that an individual belongs to a group with i opponents who have reputation G is represented as ${}_{n-1}C_i\hat{g}_t^i(1-\hat{g}_t)^{n-1-i}$. Therefore, the probability that each individual faces each of the above n situations at round t is represented as an n-dimensional vector,

$$\begin{aligned}\mathbf{s}(t) &= (s_0(t), \cdots, s_{n-1}(t)) \\ &= \Big((1-\hat{g}_t)^{n-1}, \cdots, {}_{n-1}C_i\hat{g}_t^i(1-\hat{g}_t)^{n-1-i}, \cdots, \hat{g}_t^{n-1}\Big). \end{aligned} \tag{7}$$

Using $\mathbf{s}(t)$, we represent the probability that an incumbent $\hat{\mathbf{p}}$ strategist cooperates at round t as $\hat{\epsilon}\hat{\mathbf{p}}\cdot\mathbf{s}(t)$, and the probability that an invader $\mathbf{p}$ strategist cooperates at round t as $\hat{\epsilon}\mathbf{p}\cdot\mathbf{s}(t)$.

Consequently, the frequency of individuals with reputation G among incumbent $\hat{\mathbf{p}}$ strategists at round t is

$$\hat{g}_t = \begin{cases} 1 & if \;\; t = 1 \\ \hat{\epsilon}\hat{\mathbf{p}}\cdot\mathbf{s}(t-1) & if \;\; t \geq 2. \end{cases} \tag{8}$$

Since $\hat{\mathbf{p}} = (0, \cdots, 0, 1)$, $\hat{g}_t = \hat{\epsilon}\hat{g}_{t-1}^{n-1}$ for $t \geq 2$. Considering $0 \ll \hat{\epsilon} < 1$, clearly $\lim_{t\to\infty}\hat{g}_t = \hat{g}_\infty = 0$.

Likewise, the frequency of individuals with reputation G among invader $\mathbf{p}$ strategists at round t is

$$g_t = \begin{cases} 1 & if \;\; t = 1 \\ \hat{\epsilon}\mathbf{p}\cdot\mathbf{s}(t-1) & if \;\; t \geq 2. \end{cases} \tag{9}$$

Considering $\hat{g}_\infty = 0$ and eq. (9), g_∞ is $\hat{\epsilon}p_0$ where p_0 is the first component of $\mathbf{p}$.

3.2 Expected payoff at round t

Here we derive the expected payoff at round t for incumbent $\hat{\mathbf{p}}$ strategists and for invader $\mathbf{p}$ strategists.

Recall that the payoff for an individual is determined by the probability that the focal individual cooperates and by the probability that an opponent of the focal individual cooperates (see eqs. (4) and (5)). Therefore, in order to derive the expected payoff at round t for incumbent $\hat{\mathbf{p}}$ strategists, we need to get the probability that an incumbent $\hat{\mathbf{p}}$ strategist cooperates and the probability that an opponent of the focal $\hat{\mathbf{p}}$ strategist cooperates at round t. Note that the opponent always has $\hat{\mathbf{p}}$ strategy since we consider the situation where each individual always interacts with only incumbent $\hat{\mathbf{p}}$ strategists. As we mentioned above, both the probabilities are $\hat{\epsilon}\hat{\mathbf{p}} \cdot \mathbf{s}(t)$. Consequently, the expected payoff at round t for incumbent $\hat{\mathbf{p}}$ strategists is

$$W_{(t)}(\hat{\mathbf{p}}|\hat{\mathbf{p}}) = \hat{\epsilon}b\hat{\mathbf{p}} \cdot \mathbf{s}(t) - \hat{\epsilon}c\hat{\mathbf{p}} \cdot \mathbf{s}(t) = \hat{\epsilon}\hat{\mathbf{p}} \cdot \mathbf{s}(t)(b - c). \tag{10}$$

Next, we derive the expected payoff at round t for invader $\mathbf{p}$ strategists. For this purpose, we need to get the probabilities that an invader $\mathbf{p}$ strategist cooperates at round t and that an opponent of the focal invader $\mathbf{p}$ strategist cooperates at round t. The probability that an invader $\mathbf{p}$ strategist cooperates at round t is $\hat{\epsilon}\mathbf{p} \cdot \mathbf{s}(t)$. Moreover, we consider the probability that an opponent of the focal invader $\mathbf{p}$ strategist cooperates at round t. The probability that an opponent of the focal $\mathbf{p}$ strategist faces a situation where no opponent has reputation G is $(1-g_t)(1-\hat{g}_t)^{n-2}$ and the probability that she faces a situation where all the $n-1$ opponents have reputation G is $g_t\hat{g}_t^{n-2}$. Furthermore, the probability that she faces a situation where i $(0 < i < n-1)$ opponents have reputation G is

$$(1 - g_t)_{n-2}C_i\hat{g}_t^i(1 - \hat{g}_t)^{n-2-i} + g_t \cdot_{n-2} C_{i-1}\hat{g}_t^{i-1}(1 - \hat{g}_t)^{n-1-i}. \tag{11}$$

Note that in this study, we consider the case for $n \geq 3$. The first-term indicates the case in which the focal $\mathbf{p}$ strategist in the group has reputation B and the second-term indicates the case in which the focal $\mathbf{p}$ strategist in the group has reputation G. Therefore the probability that, at round t, an opponent of the focal invader $\mathbf{p}$ strategist faces each of the n situations—there is no opponent with G reputation; there is one; $\cdots$; and there are

$n-1$ opponents with G reputation—is represented as an n-dimensional vector $\mathbf{s}'(t) = (s'_0(t), \cdots, s'_{n-1}(t))$,

$$\begin{aligned}\mathbf{s}'(t) = \Big(& (1-g_t)(1-\hat{g}_t)^{n-2}, \\ & (1-g_t)(n-2)\hat{g}_t(1-\hat{g}_t)^{n-3} + g_t(1-\hat{g}_t)^{n-2}, \\ & \cdots, \\ & (1-g_t)_{n-2}C_i\hat{g}_t^i(1-\hat{g}_t)^{n-2-i} + g_t \cdot_{n-2} C_{i-1}\hat{g}_t^{i-1}(1-\hat{g}_t)^{n-1-i}, \\ & \cdots, \\ & (1-g_t)\hat{g}_t^{n-2} + g_t(n-2)\hat{g}_t^{n-3}(1-\hat{g}_t), \\ & g_t\hat{g}_t^{n-2}\Big). \end{aligned} \tag{12}$$

Therefore, the probability that an opponent of the focal invader $\mathbf{p}$ strategist cooperates at round t is $\hat{\epsilon}\hat{\mathbf{p}}\cdot\mathbf{s}'(t)$. Note that an opponent of the focal invader $\mathbf{p}$ strategist always has $\hat{\mathbf{p}}$ strategy. Consequently, the expected payoff at round t for invader $\mathbf{p}$ strategists is

$$W_{(t)}(\mathbf{p}|\hat{\mathbf{p}}) = \hat{\epsilon}b\hat{\mathbf{p}}\cdot\mathbf{s}'(t) - \hat{\epsilon}c\mathbf{p}\cdot\mathbf{s}(t) = \hat{\epsilon}\Big(b\hat{\mathbf{p}}\cdot\mathbf{s}'(t) - c\mathbf{p}\cdot\mathbf{s}(t)\Big). \tag{13}$$

In the next section, using eqs. (10) and (13), we will derive the fitness for two types of individuals.

3.3 Fitness for incumbent $\hat{\mathbf{p}}$ *strategists and for invader* $\mathbf{p}$ *strategists*

We define the fitness for incumbent $\hat{\mathbf{p}}$ strategists, $W(\hat{\mathbf{p}}|\hat{\mathbf{p}})$, and that for invader $\mathbf{p}$ strategists, $W(\mathbf{p}|\hat{\mathbf{p}})$, as the expected value of the total payoff received by the two types of individuals, respectively, during a generation. Assuming that $W_{(t)}(\hat{\mathbf{p}}|\hat{\mathbf{p}})$ and $W_{(t)}(\mathbf{p}|\hat{\mathbf{p}})$ $(t \geq 3)$ are approximately the same as in the limit of round $t \to \infty$, $W_{(\infty)}(\hat{\mathbf{p}}|\hat{\mathbf{p}})$ and $W_{(\infty)}(\mathbf{p}|\hat{\mathbf{p}})$, $W(\hat{\mathbf{p}}|\hat{\mathbf{p}})$ and $W(\mathbf{p}|\hat{\mathbf{p}})$ are

$$W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) = W_{(1)}(\hat{\mathbf{p}}|\hat{\mathbf{p}}) + w\cdot W_{(2)}(\hat{\mathbf{p}}|\hat{\mathbf{p}}) + \frac{w^2}{1-w}W_{(\infty)}(\hat{\mathbf{p}}|\hat{\mathbf{p}}), \tag{14}$$

$$W(\mathbf{p}|\hat{\mathbf{p}}) = W_{(1)}(\mathbf{p}|\hat{\mathbf{p}}) + w\cdot W_{(2)}(\mathbf{p}|\hat{\mathbf{p}}) + \frac{w^2}{1-w}W_{(\infty)}(\mathbf{p}|\hat{\mathbf{p}}). \tag{15}$$

3.4 Evolutionary stability of $\hat{\mathbf{p}}$

Here we show that the inequality $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}|\hat{\mathbf{p}})$ holds for all $\mathbf{p} \neq \hat{\mathbf{p}}$.

3.4.1 Evolutionary stability of $\hat{\mathbf{p}}$ against invasion by $\mathbf{p}$ with $p_{n-1} = 1$

For this purpose, we first show that $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}|\hat{\mathbf{p}})$ for all $\mathbf{p}$ ($\neq \hat{\mathbf{p}}$) with $p_{n-1} = 1$, hereafter called $\mathbf{p}^C = (p_0^C, \cdots, p_{n-2}^C, 1)$.

Substituting eq. (10) into eq. (14), the fitness for incumbent $\hat{\mathbf{p}}$ strategists is

$$\begin{aligned} W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) &= \hat{\epsilon}(b-c) + w\hat{\epsilon}\hat{\mathbf{p}}\cdot\mathbf{s}(2)(b-c) + \frac{w^2}{1-w}\hat{\epsilon}\hat{\mathbf{p}}\cdot\mathbf{s}(\infty)(b-c) \\ &= \hat{\epsilon}(b-c) + w\hat{\epsilon}\hat{\mathbf{p}}\cdot\mathbf{s}(2)(b-c). \end{aligned} \tag{16}$$

Note that since $\hat{g}(\infty) = 0$, $\hat{\mathbf{p}}\cdot\mathbf{s}(\infty) = 0$.

On the other hand, substituting eq. (13) into eq. (15), the fitness for invader $\mathbf{p}^C$ strategists is

$$\begin{aligned} W(\mathbf{p}^C|\hat{\mathbf{p}}) &= \hat{\epsilon}(b-c) + w\hat{\epsilon}[b\hat{\mathbf{p}}\cdot\mathbf{s}'(2) - c\mathbf{p}^C\cdot\mathbf{s}(2)] \\ &\quad + \frac{w^2}{1-w}\hat{\epsilon}[b\hat{\mathbf{p}}\cdot\mathbf{s}'(\infty) - c\mathbf{p}^C\cdot\mathbf{s}(\infty)] \\ &= \hat{\epsilon}(b-c) + w\hat{\epsilon}[b\mathbf{p}\cdot\mathbf{s}'(2) - c\mathbf{p}^C\cdot\mathbf{s}(2)] - \frac{w^2}{1-w}\hat{\epsilon}cp_0^C. \end{aligned} \tag{17}$$

Note that since we are considering the case for $n \geq 3$, $s'_{n-1}(\infty) = 0$ and so $\hat{\mathbf{p}}\cdot\mathbf{s}'(\infty) = 0$. Furthermore, $\mathbf{s}(\infty) = (1, 0, \cdots, 0)$ because $\hat{g}(\infty) = 0$. Therefore, $\mathbf{p}^C\cdot\mathbf{s}(\infty) = p_0^C$.

Since both types of individuals intend to cooperate at the first round, $\hat{g}_2 = g_2$, $\mathbf{s}(2) = \mathbf{s}'(2)$. Considering $\mathbf{s}(2) = \mathbf{s}'(2)$, the difference between the fitness for $\hat{\mathbf{p}}$ and $\mathbf{p}^C$ strategists is

$$W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) - W(\mathbf{p}^C|\hat{\mathbf{p}}) = w\hat{\epsilon}c\mathbf{s}(2)\cdot(\mathbf{p}^C - \hat{\mathbf{p}}) + \frac{w^2}{1-w}\hat{\epsilon}cp_0^C. \tag{18}$$

The difference (eq. (18)) is 0 if and only if $\mathbf{p}^C = \hat{\mathbf{p}}$, i.e., $\mathbf{p}^C = (0, \cdots, 0, 1)$; otherwise the difference is positive [4]. Consequently,

$$W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}^C|\hat{\mathbf{p}}) \quad (for\ all\ \mathbf{p}^C \neq \hat{\mathbf{p}}), \tag{19}$$

where $\mathbf{p}^C$ is a strategy $\mathbf{p}$ with $p_{n-1} = 1$.

[4] Note that since all the components of $\mathbf{s}(2)$ are positive and those of $\mathbf{p}^C - \hat{\mathbf{p}}$ are non-negative, these two vectors are not orthogonal.

3.4.2 Evolutionary stability of $\hat{\mathbf{p}}$ against invasion by $\mathbf{p}$ with $p_{n-1} = 0$

Next, we show that the condition that the inequality $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}|\hat{\mathbf{p}})$ holds for all $\mathbf{p}$ with $p_{n-1} = 0$, hereafter called $\mathbf{p}^D = (p_0^D, \cdots, p_{n-2}^D, 0)$. To show this, we first show that the fitness for the $\mathbf{p}^0 = (0, \cdots, 0)$ strategy which is one of the $\mathbf{p}^D$ strategies, is the largest among the fitness for the $\mathbf{p}^D$ strategies, i.e., $W(\mathbf{p}^0|\hat{\mathbf{p}}) > W(\mathbf{p}^D|\hat{\mathbf{p}})$ for all $\mathbf{p}^D \neq \mathbf{p}^0$.

Substituting eq. (13) into eq. (15), the fitness for invader $\mathbf{p}^0$ strategists is

$$W(\mathbf{p}^0|\hat{\mathbf{p}}) = \hat{\epsilon}b. \tag{20}$$

Note that since $\mathbf{p}^0$ strategists never cooperate and so $g_t = 0$ for $t \geq 2$, incumbent $\hat{\mathbf{p}}$ strategists never cooperate with them.

On the other hand, by $p_{n-1}^D = 0$ and eq. (9), the frequency of individuals with reputation G among $\mathbf{p}^D$ strategists is as follows: $g_2 = 0$ and $g(\infty) = \hat{\epsilon}p_0^D$. Therefore, substituting eq. (13) into eq. (15), the fitness for invader $\mathbf{p}^D$ strategists is

$$\begin{aligned} W(\mathbf{p}^D|\hat{\mathbf{p}}) &= \hat{\epsilon}b - w\hat{\epsilon}c\mathbf{p}^D \cdot \mathbf{s}(2) - \frac{w^2}{1-w}\hat{\epsilon}c\mathbf{p}^D \cdot \mathbf{s}(\infty) \\ &= \hat{\epsilon}b - w\hat{\epsilon}c\mathbf{p}^D \cdot \mathbf{s}(2) - \frac{w^2}{1-w}\hat{\epsilon}cp_0^D. \end{aligned} \tag{21}$$

Note that $\mathbf{s}(\infty) = (1, 0, \cdots, 0)$.

The difference between the fitness for invader $\mathbf{p}^0$ and $\mathbf{p}^D$ strategists is

$$W(\mathbf{p}^0|\hat{\mathbf{p}}) - W(\mathbf{p}^D|\hat{\mathbf{p}}) = w\hat{\epsilon}c\mathbf{p}^D \cdot \mathbf{s}(2) + \frac{w^2}{1-w}\hat{\epsilon}cp_0^D. \tag{22}$$

Since $s_i(2) > 0$ for all i $(0 \leq i \leq n-1)$, the difference (eq. (22)) is 0 if and only if $\mathbf{p}^D = \mathbf{p}^0$, i.e., $\mathbf{p}^D = (0, \cdots, 0)$; otherwise the difference is positive [5]. Therefore, $W(\mathbf{p}^0|\hat{\mathbf{p}}) > W(\mathbf{p}^D|\hat{\mathbf{p}})$ for all $\mathbf{p}^D \neq \mathbf{p}^0$.

Hence, $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}^D|\hat{\mathbf{p}})$ for all $\mathbf{p}^D$ if and only if $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}^0|\hat{\mathbf{p}})$. Comparing eq. (20) with eq. (16), we find that the condition for $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}^0|\hat{\mathbf{p}})$ is

$$W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}^0|\hat{\mathbf{p}})$$

[5] Note that since all the components of $\mathbf{s}(2)$ are positive and those of $\mathbf{p}^D$ are non-negative, these two vectors are not orthogonal.

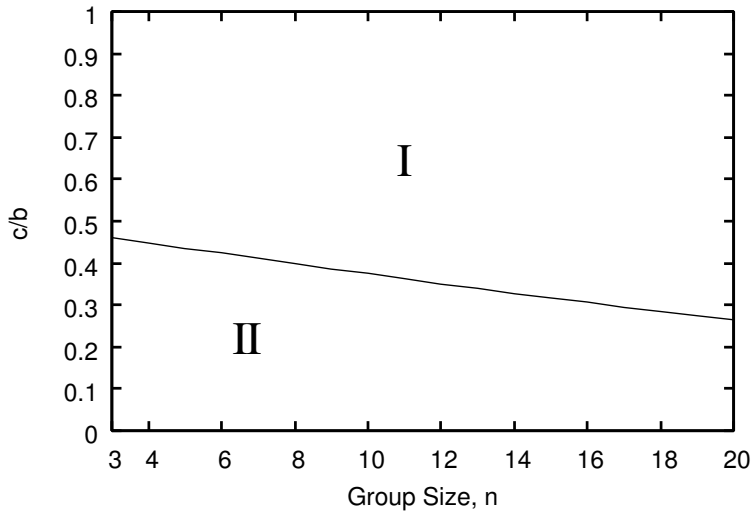

Fig. 1. Regions in $(n,\ c/b)$ space, where $\hat{\mathbf{p}}$ is an ESS, given analytically with the approximation ($\hat{\epsilon} = 0.95$ and $w = 0.95$). Region I: $\hat{\mathbf{p}}$ can not resist invasion by $\mathbf{p}^D$. Region II: $\hat{\mathbf{p}}$ is an ESS.

$$\hat{\epsilon}(b-c) + w\hat{\epsilon}^n(b-c) > \hat{\epsilon}b$$
$$c/b < \frac{w\hat{\epsilon}^{n-1}}{1+w\hat{\epsilon}^{n-1}}. \tag{23}$$

Therefore, $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}^D|\hat{\mathbf{p}})$ for all $\mathbf{p}^D$ if and only if $c/b < w\hat{\epsilon}^{n-1}/(1 + w\hat{\epsilon}^{n-1})$.

3.4.3 Evolutionary stability of $\hat{\mathbf{p}}$ *against invasion by all the* $\mathbf{p}$ *strategists*

We have shown that $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}|\hat{\mathbf{p}})$ for all $\mathbf{p}$ $(\neq \hat{\mathbf{p}})$ with $p_{n-1} = 1$ and that $W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}|\hat{\mathbf{p}})$ for all $\mathbf{p}$ with $p_{n-1} = 0$ if and only if $c/b < w\hat{\epsilon}^{n-1}/(1+ w\hat{\epsilon}^{n-1})$.

Consequently,

$$W(\hat{\mathbf{p}}|\hat{\mathbf{p}}) > W(\mathbf{p}|\hat{\mathbf{p}}) \quad (for\ all\ \mathbf{p} \neq \hat{\mathbf{p}}), \tag{24}$$

if and only if $c/b < w\hat{\epsilon}^{n-1}/(1 + w\hat{\epsilon}^{n-1})$. In other words, under image scoring, the strategy $\hat{\mathbf{p}}$ is an ESS if the cost-to-benefit ratio of cooperation is sufficiently small. The condition of the cost-to-benefit ratio is illustrated in Fig. 1. As shown in the figure, the condition becomes more restrictive as group size increases, but not so drastically.

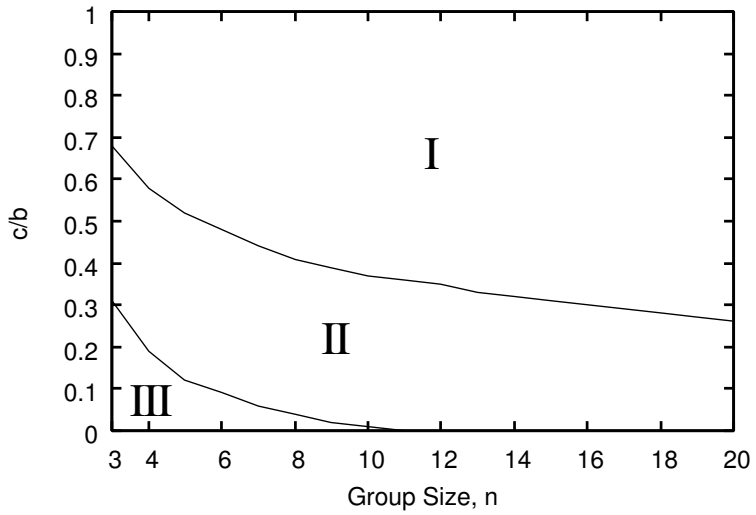

Fig. 2. Regions in $(n,\ c/b)$ space, where $\hat{\mathbf{p}}$ is an ESS, given numerically without the approximation ($\hat{\epsilon} = 0.95$ and $w = 0.95$). Region I: $\hat{\mathbf{p}}$ can not resist invasion by $\mathbf{p}^D$. Region II: $\hat{\mathbf{p}}$ is an ESS. Region III: $\hat{\mathbf{p}}$ can not resist invasion by $\mathbf{p}^C$.

4 Evolutionary stability of indirect reciprocity (numerical investigation)

So far, we have investigated the evolutionary stability of the indirectly reciprocal strategy $\hat{\mathbf{p}}$ analytically with the approximation given in eqs (14)-(15). Here, we investigate the evolutionary stability using numerical calculation without the approximation.

We illustrate the result of the numerical calculation in Fig. 2. As shown in the figure, between small and large cost-to-benefit ratio of cooperation, there exist a range of the medium ratio where the indirectly reciprocal strategy $\hat{\mathbf{p}}$ is an ESS. If the cost-to-benefit ratio is large, $\hat{\mathbf{p}}$ can not resist invasion by $\mathbf{p}^D$, and, if the ratio is extremely small, $\hat{\mathbf{p}}$ can not resist invasion by $\mathbf{p}^C$. Under the medium cost-to-benefit ratio of cooperation, $\hat{\mathbf{p}}$ is an ESS. In other words, the conclusion of the analyses that $\hat{\mathbf{p}}$ can be an ESS under image scoring in $n(> 2)$-person game does not change whether the approximation is used or not.

Furthermore, comparing Fig. 2 with Fig. 1, we find that the approximation given in eqs (14)-(15), which makes possible the analytical investigation in the previous section, is valid especially for large group size. The validity of the approximation for large group is crucial for the investigation of evolutionary stability, because, for large group size, we have to check the stability of $\hat{\mathbf{p}}$ against a huge number of strategies in order to investigate the evolutionary stability of $\hat{\mathbf{p}}$ and it requires huge computational time. For instance, if group size, $n = 20$, we have to check the stability against about million strategies.

5 Discussion

In this paper, we have analyzed the evolutionary stability of indirectly reciprocal strategy, DIS ($\hat{\mathbf{p}}$ strategy) which cooperates only when all the opponents in the group have good reputations, under image scoring. Under image scoring which reflects only first-order-information, cooperations (defections) are judged to be good (bad) unconditionally. In particular, we have focused on the n (> 2) person case where n persons take part in a single group, and on the case where there is implementation error. The analysis has shown that, in the n-person case, DIS can be an ESS even under image scoring, which reflects only first-order information, whereas DIS can never be an ESS in the two-person case (e.g. Panchanathan & Boyd, 2003; Ohtsuki & Iwasa, 2004; Brandt & Sigmund, 2004; Takahashi & Mashima, 2006). In other words, in the n (> 2) person case, first-order information is sufficient for indirect reciprocity to be maintained stably.

Specifically, in the n (> 2) person case, DIS is an ESS among all the possible strategies that decide their own action based on the number of opponents whose reputation is good, within a range of the cost-to-benefit ratio of cooperation, even under image scoring. If the cost-to-benefit ratio is large, DIS can not resist invasion by defective strategies and, if the ratio is extremely small, DIS can not resist invasion by cooperative strategies. Under the medium cost-to-benefit ratio of cooperation, DIS is an ESS. Moreover, the condition for DIS to be an ESS becomes more restrictive as group size, n, increases, but not so drastically. However, it is important to note that image scoring in the n (> 2) person case does not satisfy Ohtsuki & Iwasa's criterion for the leading eight (Ohtsuki & Iwasa, 2004) under which a strategy that attains high level cooperation is an ESS, because DIS cannot attain high level cooperation under image scoring in the presence of error. Under image scoring, the frequency of individuals who have good reputation decreases over time in response to error defections, and so they come not to cooperate with each other.

The difference in the evolutionary stability of DIS under image scoring between the two-person case and the n (> 2) person case is attributed to the ability to resist invasion by $ALLC$. Consider the situation in which DIS strategists are incumbent and a few $ALLC$ strategists invade the population. In the two-person case, a DIS strategist does not cooperate with individuals who defected because of error, so her reputation gets worse. As a result, DIS strategists hurt each other. On the contrary, an $ALLC$ strategist is not drawn into the chain of retaliative defections because she always intends to cooperate, and thus her reputation does not get worse. Therefore, the fitness of $ALLC$ exceeds that of DIS and so DIS cannot resist invasion by $ALLC$. On the other hand, in the n (> 2) person case, an invader $ALLC$ strategist mostly belongs to a group with two or more incumbent DIS strategists. In this group,

the DIS strategists defect in response to another DIS strategist's bad reputation. That is, an invader $ALLC$ strategist cannot avoid being drawn into the chain of retaliative defections in the n (> 2) person case. Therefore, the fitness of $ALLC$ is less than that of DIS because $ALLC$ strategists intend to cooperate even in the chain of the retaliative defections, and so DIS strategists can resist invasion by $ALLC$. Consequently, DIS is an ESS under image scoring in the n (> 2) person case, whereas DIS is not in the two-person case.

Acknowledgements This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (S), 2005–2009, 17103002, and for JSPS Fellows, 2006-2007, 183845.

References

Alexander, R. D. 1987. *The Biology of Moral Systems*. Aldine de Gruyter, New York.

Bolton, G. E., Katok, E., & Ockenfels, A. 2005. Cooperation among strangers with limited information about reputation. *J. Public. Econ.* **89**, 1457-1468.

Boyd, R., & Richerson, P. J. 1988. The Evolution of Reciprocity in Sizable Groups. *J. Theor. Biol.* **132**, 337-356.

Brandt, H., & Sigmund, K. 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. *J. Theor. Biol.* **231**, 475-486.

Brandt, H. & Sigmund, K. 2005. Indirect reciprocity, image scoring, and moral hazard. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2666-2670.

Eriksson, A. & Lindgren, K. 2005. Cooperation driven by mutations in multiperson Prisoner's Dilemma. *J. Theor. Biol.* **232**, 399-409.

Fishman, M. A. 2003. Indirect reciprocity among imperfect individuals. *J. Theor. Biol.* **225**, 285-292.

Joshi, N. V. 1987. Evolution of cooperation by reciprocation within structured demes. *J. Genet.* **66**, 69-84.

Leimar, O. & Hammerstein, P. 2001. Evolution of cooperation through indirect reciprocity. *Proc. R. Soc. Lond.* **B 268**, 745-753.

Lindgren, K & Johansson, J. 2001. Coevolution of strategies in n-person Prisoner's Dilemma. *In Evolutionary Dynamics: Exploring the Interplay of Selection, Accident, Neutrality, and Function.* Edited by James P. Crutchfield and Peter Schuster, Santa Fe Institute Studies in the Sciences of Complexity Series.

Lotem, A., Fishman. M. A. & Stone, L. 1999. Evolution of cooperation between individuals. *Nature*. **400**, 226-227.

Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H. J. 2001. Cooperation through indirect reciprocity: image scoring or standing strategy?. *Proc. R. Soc. Lond.* **B 268**, 2495-2501.

Milinski, M., Semmann, D. & Krambeck, H. J. 2002b. Reputation helps solve the Tragedy of the commons. *Nature*. **415**, 424-426.

Mohtashemi, M. & Mui, L. 2003. Evolution of indirect reciprocity by social information: the role of trust and reputation in evolution of altruism. *J. Theor. Biol*. **223**, 523-531.

Molander, P. 1992. The Prevalence of Free Riding. *J. Conflict. Resolt*. **36**, 756-771.

Nowak, M. A. & Sigmund, K. 1998a. Evolution of indirect reciprocity by image scoring. *Nature*. **393**, 573-577.

Nowak, M. A., & Sigmund, K. 1998b. The dynamics of indirect reciprocity. *J. Theor. Biol*. **194**, 561-574.

Nowak, M. A. & Sigmund, K. 2005. Evolution of indirect reciprocity. *Nature*. **437**, 1291-1298.

Ohtsuki, H. & Iwasa, Y. 2004. How should we define goodness? – reputation dynamics in indirect reciprocity. *J. Theor. Biol*. **231**, 107-120.

Ohtsuki, H. & Iwasa, Y. 2005. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol*. **239**, 435-444.

Panchanathan, K. & Boyd, R. 2003. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J. Theor. Biol*. **224**, 115-126.

Panchanathan, K. & Boyd, R. 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*. **432**, 499-502.

Sugden, R. 1986. *The Evolution of Rights, Co-operation and Welfare*. Oxford: Blackwell.

Suzuki, S. & Akiyama, E. 2005. Reputation and the evolution of cooperation in sizable groups. *Proc. R. Soc. Lond*. **B 272**, 1373-1377.

Suzuki, S. & Akiyama, E. submitted. Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity.

Takahashi, N. & Mashima, R. 2006. The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. *J. Theor. Biol*. forthcoming.

Wedekind, C. & Milinski, M. 2000. Cooperation Through Image Scoring in Humans. *Science*. **288**, 850-852.