A second-order perturbation theory for the continuous model of indirect reciprocity

Sanghun Lee

Department of Physics, Pukyong National University, Busan 48513, Korea

Yohsuke Murase

RIKEN Center for Computational Science, Kobe, Hyogo 650-0047, Japan

Seung Ki Baek

Department of Scientific Computing, Pukyong National University, Busan 48513, Korea

Abstract

Reputation is one of key mechanisms to maintain human cooperation, but its analysis gets complicated if we consider the possibility that reputation does not reach consensus because of erroneous assessment. The difficulty is alleviated if we assume that reputation and cooperation do not take binary values but have continuous spectra so that disagreement over reputation can be analysed through a perturbation theory. In this work, we carry out the analysis by expanding the dynamics of reputation to the second order of perturbation under the assumption that everyone initially cooperates with good reputation. The second-order theory clarifies the difference between Image Scoring and Simple Standing in that punishment for defection against a well-reputed player should be regarded as good for maintaining cooperation. Moreover, comparison among the leading eight shows that the stabilizing effect of justified punishment weakens if cooperation between two ill-reputed players is regarded as bad. Our analysis thus explains how Simple Standing achieves a high level of stability by permitting justified punishment and also by disregarding irrelevant information in assessing cooperation. This observation suggests which factors affect the stability of a social norm when reputation can be perturbed by noise.

Email address: seungki@pknu.ac.kr (Seung Ki Baek)
1. Introduction

The power and the instability of reputation have attracted interest among researchers in the field of social evolution\cite{1}. Reputation strongly affects our behaviour from early childhood\cite{2}, but it can turn to a capricious tyrant: Sometimes a small mistake ruins it without deserving, and it may be unrecoverable nowadays when digital footprints last forever. The question is how to make a stable reputation system that recovers from erroneous assessment.

The dynamics of reputation was first analysed in mathematical terms by considering a norm called Image Scoring\cite{3}. It is a first-order norm in the sense that it assigns reputation to an individual depending on what she did to her co-player\cite{4}. However, by ignoring the co-player’s reputation\cite{5}, it falls outside of the leading eight, the set of cooperative norms that are evolutionarily stable against every behavioural mutant\cite{6, 7} (Table\ref{table:1}). The key point is that one must be allowed to refuse to cooperate toward an ill-reputed co-player without risking her own reputation: Otherwise, well-intentioned punishment will not be distinguished from malicious defection, and conditional cooperators cannot thrive in such an environment. Indeed, according to the leading eight, a well-reputed player’s defection against an ill-reputed co-player should be regarded as good so as to secure cooperation at the societal level. We believe that this property should generally be true even beyond the binary-reputation system\cite{8}.

How to ensure stable cooperation in the presence of error and private reputation rules is still under active investigation\cite{9, 10, 11, 12, 13, 14, 15}. In our previous work\cite{16}, we proposed to regard reputation and cooperation as continuous variables to calculate the effects of different assessments in a perturbative way. This continuum approach offered a simple understanding of the reason that some of the leading eight are vulnerable to error in reputation. In addition, by assuming small difference between the resident and mutant norms, we derived a threshold for the benefit-to-cost ratio of cooperation to suppress mutants\cite{16}, replacing an earlier prediction relating the threshold to the probability of observation\cite{3, 4, 17}. However, by taking into account only linear-order perturbation, our previous work failed to address the difference between first- and second-order norms because the simultane-
Table 1: Leading eight and Image Scoring (IS). We denote cooperation and defection as $C$ and $D$, respectively, and a player’s reputation as either good (1) or bad (0). By $\alpha_{uXv}$, therefore, we mean the reputation assigned to a player who had reputation $u$ and did $X \in \{C, D\}$ to another player with reputation $v$. The behavioural rule $\beta_{uv}$ prescribes an action between $C$ and $D$ when the focal player has reputation $u$ and the co-player has reputation $v$.

|      | $\alpha_{1C1}$ | $\alpha_{1D1}$ | $\alpha_{1C0}$ | $\alpha_{1D0}$ | $\alpha_{0C1}$ | $\alpha_{0D1}$ | $\alpha_{0C0}$ | $\alpha_{0D0}$ | $\beta_{11}$ | $\beta_{10}$ | $\beta_{01}$ | $\beta_{00}$ |
|------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-------------|-------------|-------------|-------------|
| L1   | 1              | 0              | 1              | 1              | 0              | 1              | 0              | 0              | C           | D           | C           | C           |
| L2   | 1              | 0              | 0              | 1              | 1              | 0              | 0              | 1              | C           | D           | C           | C           |
| L3   | 1              | 0              | 1              | 1              | 0              | 1              | 1              | 0              | C           | D           | C           | D           |
| L4   | 1              | 0              | 1              | 1              | 0              | 1              | 0              | 0              | C           | D           | C           | D           |
| L5   | 1              | 0              | 0              | 1              | 1              | 0              | 1              | 1              | C           | D           | C           | D           |
| L6   | 1              | 0              | 0              | 1              | 1              | 0              | 0              | 1              | C           | D           | C           | D           |
| L7   | 1              | 0              | 1              | 1              | 0              | 1              | 0              | 0              | C           | D           | C           | D           |
| L8   | 1              | 0              | 0              | 1              | 1              | 0              | 0              | 0              | C           | D           | C           | D           |
| IS   | 1              | 0              | 1              | 0              | 1              | 0              | 1              | 0              | C           | D           | C           | D           |

ous action of defection and bad reputation appears as a second-order effect if everyone initially cooperates with good reputation.

In this work, we wish to fill this gap by extending the perturbation theory to the second order. Our analysis identifies second-order effects such as justified punishment on the stability of the reputation system when perturbation is caused by erroneous assessment.

2. Analysis

The basic dynamical process goes as follows: At every time step, we pick up a random pair of players, say, $i$ and $j$, as a donor and a recipient, respectively, from a large population of size $N \gg 1$. The conventional setting is that they play the one-shot donation game so that the donor makes a binary decision on whether to donate $b$ to the recipient by paying $c$, where $b$ and $c$ represent the benefit and the cost of cooperation, respectively, with $b > c > 0$. In our continuous version of the donation game, the donor $i$ chooses the level of donation, $\beta_i \in [0, 1]$, so that $b\beta_i$ is donated to $j$ at the cost of $c\beta_i$. Cooperation and defection thus correspond to $\beta_i = 1$ and 0, respectively.

Players in the population observe the interaction with probability $q$. Let $k$ be one of the observers. We introduce $m_{ki}$ as a continuous variable between $0$
(bad) and 1 (good), for describing $i$'s reputation according to $k$'s assessment rule $\alpha_k = \alpha_k(m_{ki}, \beta_i, m_{kj})$. Likewise, the donation level $\beta_i$ depends on how the donor assesses herself as well as the recipient, i.e., $\beta_i = \beta_i(m_{ii}, m_{ij})$. As time goes by, the new assessment replaces the older one, so the averaged dynamics of $m_{ki}$ in the continuous-time limit can be written as follows [15, 16]:

$$\frac{d}{dt}m_{ki} = -qm_{ki} + \frac{q}{N-1} \sum_{j \neq i} \alpha_k [m_{ki}, \beta_i(m_{ii}, m_{ij}), m_{kj}], \quad (1)$$

where $q$ is the probability of observation. We assume that implementation error and perception error occur at a low rate, so that error plays the role of perturbation to the initial condition without affecting the governing equation itself.

Let us assume that the society has adopted a common social norm $(\alpha, \beta)$, one of whose stationary fixed points is a fully cooperative initial state with $m_{ij} = 1$ for every pair of $i$ and $j$. In other words, we consider norms that satisfy

$$\alpha(1,1,1) = \beta(1,1) = 1. \quad (2)$$

Error perturbs the players’ reputations from this fixed point, and we are interested in how small perturbation $\epsilon_{ki} \equiv 1 - m_{ki}$ grows over time. By expanding Eq. (1) to the first order in $\epsilon_{ki}$’s, we obtain the following equation:

$$\frac{d}{dt}\epsilon_{ki} \approx -q(1 - A_x)\epsilon_{ki} + qA_yB_x\epsilon_{ii} + \frac{q}{N-1} \sum_{j \neq i} [A_yB_y\epsilon_{ij} + A_z\epsilon_{kj}], \quad (3)$$

where

$$A_x \equiv \partial_x \alpha(x, y, z)|_{(1,1,1)} \quad (4a)$$

$$A_y \equiv \partial_y \alpha(x, y, z)|_{(1,1,1)} \quad (4b)$$

$$A_z \equiv \partial_z \alpha(x, y, z)|_{(1,1,1)} \quad (4c)$$

$$B_x \equiv \partial_x \beta(x, y)|_{(1,1)} \quad (4d)$$

$$B_y \equiv \partial_y \beta(x, y)|_{(1,1)}. \quad (4e)$$

To make the story more concrete, we can construct a continuous version of Simple Standing [18], denoted by L3 in Table III, by using the bilinear and
Table 2: Continuous versions of the leading eight and Image Scoring. The assessment rule \( \alpha(x, y, z) \) is obtained by applying the trilinear interpolation to \( \alpha_{xyz} \)'s in Table 1, where \( C \) and \( D \) correspond to 1 and 0, respectively. Likewise, the behavioural rule \( \beta(x, y) \) results from the bilinear interpolation applied to \( \beta_{xy} \)'s. We note that L1 has been nicknamed Contrite Tit-for-Tat in the context of direct reciprocity [18, 19].

| Norm               | \( \alpha(x, y, z) \)                  | \( \beta(x, y) \)                  |
|--------------------|----------------------------------------|------------------------------------|
| L1                 | \( x + y - xy - xz + xyz \)            | \( -x + xy + 1 \)                  |
| L2 (Consistent Standing) | \( x + y - 2xy - xz + 2xyz \)        | \( -x + xy + 1 \)                  |
| L3 (Simple Standing)      | \( yz - z + 1 \)                     | \( y \)                            |
| L4                 | \( -y - z + xy + 2yz - xyz + 1 \)      | \( y \)                            |
| L5                 | \( -z - xy + yz + xyz + 1 \)           | \( y \)                            |
| L6 (Stern Judging)      | \( -y - z + 2yz + 1 \)                | \( y \)                            |
| L7 (Staying)         | \( x - xz + yz \)                     | \( y \)                            |
| L8 (Judging)         | \( x - xy - xz + yz + xyz \)          | \( y \)                            |
| Image Scoring       |                                        | \( y \)                            |

The trilinear interpolation methods as follows:

\[
\alpha_{SS}(x, y, z) = yz - z + 1 \quad (5a) \\
\beta_{SS}(x, y) = y. \quad (5b)
\]

Table 2 shows the full list of interpolated expressions for the leading eight and Image Scoring. If we look at Table 3 all the leading eight have \( A_x = B_x = 0 \) and \( A_y = B_y = 1 \) in this linear description. The only difference among the leading eight lies in \( A_z \): That is, L1, L3, L4, and L7 have \( A_z = 0 \), whereas L2, L5, L6, and L8 have \( A_z = 1 \). As for Image Scoring, the expression is even simpler:

\[
\alpha_{IS} = \beta_{IS} = y. \quad (6)
\]

Note that the linear-order description for Simple Standing and Image Scoring is given as

\[
(A_x, A_y, A_z, B_x, B_y) = (0, 1, 0, 0, 1) \quad (7)
\]

in common. Their difference is manifested in the second-order description because \( A_{yz} = 1 \) for Simple Standing, whereas \( A_{yz} = 0 \) for Image Scoring as shown in Table 4 where \( A_{\mu\nu} \equiv \partial^2 \alpha/\partial \mu \partial \nu \mid_{(1,1)} \) and \( B_{\mu\nu} \equiv \partial^2 \beta/\partial \mu \partial \nu \mid_{(1,1)} \).

Equation (3) can be expressed as a linear-algebraic equation for an \( N^2 \)-dimensional vector \( \vec{V} = (\epsilon_{11}, \ldots, \epsilon_{NN}) \). The \( N^2 \times N^2 \) matrix acted on \( \vec{V} \) has
Table 3: First-order derivatives of the continuous leading eight and Image Scoring at 
\((x, y, z) = (1, 1, 1)\). Note that their differences lie only in \(A_z\).

| Norm       | \(A_x\) | \(A_y\) | \(A_z\) | \(B_x\) | \(B_y\) |
|------------|--------|--------|--------|--------|--------|
| L1         | 0      | 1      | 0      | 0      | 1      |
| L2         | 0      | 1      | 1      | 0      | 1      |
| L3         | 0      | 1      | 0      | 0      | 1      |
| L4         | 0      | 1      | 0      | 0      | 1      |
| L5         | 0      | 1      | 1      | 0      | 1      |
| L6         | 0      | 1      | 1      | 0      | 1      |
| L7         | 0      | 1      | 0      | 0      | 1      |
| L8         | 0      | 1      | 1      | 0      | 1      |
| Image Scoring | 0      | 1      | 0      | 0      | 1      |

the largest eigenvalue in the following form \[16\]:

\[
\Lambda_1 = q \left[ -1 + A_x + A_z + A_y(B_x + B_y) \right],
\]

and the corresponding eigenvector is

\[
\vec{V}_1 = (1, 1, \ldots, 1)
\]

up to a proportionality constant. If \(\Lambda_1 > 0\), the fixed point in Eq. (2) is
unstable, which is the case of L2, L5, L6, and L8 because they have \(A_z = 1\). For the others norms of the leading eight as well as for Image Scoring, this linear-order analysis leaves stability indeterminate by having \(\Lambda_1 = 0\). From our viewpoint, the important point is that Simple Standing is not
distinguished from Image Scoring because they have exactly the same first
derivatives \[Eq. (7)\].

To proceed, we take into account second-order terms to write down the
following equation (see Appendix [Appendix A]):

\[
\frac{d\epsilon_{ki}}{dt} = -q\epsilon_{ki} - \frac{q}{N-1} \sum_{j \neq i} \left[ -A_x\epsilon_{ki} - A_y(B_x\epsilon_{ii} + B_y\epsilon_{ij}) - A_z\epsilon_{kj} \right]
\]

\[
- \frac{q}{N-1} \sum_{j \neq i} \left[ A_y \left( \frac{1}{2}B_{xx}\epsilon_{ii}^2 + B_{xy}\epsilon_{ii}\epsilon_{ij} + \frac{1}{2}B_{yy}\epsilon_{ij}^2 \right) \right]
\]

\[
+ \frac{1}{2}A_{xx}\epsilon_{ki}^2 + \frac{1}{2}A_{yy}(B_x\epsilon_{ii} + B_y\epsilon_{ij})^2 + \frac{1}{2}A_{zz}\epsilon_{kj}^2
\]

\[
+ A_{xy}\epsilon_{ki}(B_x\epsilon_{ii} + B_y\epsilon_{ij}) + A_{yz}(B_x\epsilon_{ii} + B_y\epsilon_{ij})\epsilon_{kj} + A_{xz}\epsilon_{ki}\epsilon_{kj} \right] + \ldots
\]
Table 4: Second-order derivatives of the continuous leading eight and Image Scoring at 
(x, y, z) = (1, 1, 1). Note that their differences lie only in $A_{zx}$, $A_{yz}$, and $B_{xy}$.

| Norm          | $A_{xx}$ | $A_{xy}$ | $A_{xz}$ | $A_{yy}$ | $A_{yz}$ | $A_{zz}$ | $B_{xx}$ | $B_{xy}$ | $B_{yy}$ |
|---------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| L1            | 0       | 0       | 0       | 0       | 1       | 0       | 0       | 1       | 0       |
| L2            | 0       | 0       | 1       | 0       | 2       | 0       | 0       | 1       | 0       |
| L3            | 0       | 0       | 0       | 0       | 1       | 0       | 0       | 0       | 0       |
| L4            | 0       | 0       | -1      | 0       | 1       | 0       | 0       | 0       | 0       |
| L5            | 0       | 0       | 1       | 0       | 2       | 0       | 0       | 0       | 0       |
| L6            | 0       | 0       | 0       | 0       | 2       | 0       | 0       | 0       | 0       |
| L7            | 0       | 0       | -1      | 0       | 1       | 0       | 0       | 0       | 0       |
| L8            | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 0       |
| Image Scoring | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 0       |

We wish to reduce the $N^2$-dimensional dynamics into a one-dimensional 
one along the principal eigenvector $\vec{V}_1 = (\epsilon, \epsilon, \ldots, \epsilon)$ [Eq. (9)]. To argue the 
validity of this approximation, let us write $\epsilon_{ki} = \epsilon + c_{ki}$, where $|c_{ki}| \ll \epsilon$. We 
then decompose Eq. (10) into two parts, i.e., one for $\epsilon$ and the other for $c_{ki}$. The former is written as

$$
\frac{d\epsilon}{dt} \approx \Lambda_1 \epsilon - q \left[ \frac{1}{2} A_{xx} + A_{zx} + \frac{1}{2} A_{zz} + (A_{xy} + A_{yz}) (B_x + B_y) \\
+ \frac{1}{2} A_{yy} (B_x + B_y)^2 + A_y \left( \frac{1}{2} B_{xx} + B_{xy} + \frac{1}{2} B_{yy} \right) \right] \epsilon^2, \quad (11)
$$

and the latter obeys the following dynamics:

$$
\frac{dc_{ki}}{dt} \approx q \{(-1 + A_x) c_{ki} + A_y B_x c_{ii} \} + \frac{q}{N-1} \sum_{j \neq i} (A_y B_y c_{ij} + A_z c_{kj}) \\
- \frac{q \epsilon}{N-1} \sum_{j \neq i} [A_y (B_{xx} c_{ii} + B_{xy} (c_{ii} + c_{ij}) + B_{yy} c_{ij})] \\
+ A_{xx} c_{ki} + A_{yy} (B_x c_{ii} + B_y c_{ij}) (B_x + B_y) + A_{zz} c_{kj} \\
+ (A_{xy} + A_{yz}) (B_x c_{ii} + B_y c_{ij}) + (A_{xy} c_{ki} + A_{yz} c_{kj}) (B_x + B_y) \\
+ A_{zx} (c_{ki} + c_{kj})], \quad (12)
$$

where $\epsilon$ is regarded as a constant. For the norms with $\Lambda_1 = 0$, i.e., L1, L3, L4, L7, and Image Scoring, the largest eigenvalue of Eq. (12) is either negative or zero (see Appendix B). Specifically, it is $-2q\epsilon$ for L1 and L3.
Table 5: Stability of the initial cooperative state with $\epsilon = 0$ under each of the leading eight and Image Scoring in the continuous model.

| Norm                | Dynamics                                      | Stability of $\epsilon = 0$ |
|---------------------|-----------------------------------------------|------------------------------|
| L1                  | $\frac{d\epsilon}{dt} = -2q\epsilon^2 + \ldots$ | Stable                       |
| L2 (Consistent Standing) | $\frac{d\epsilon}{dt} = q\epsilon + \ldots$      | Unstable                     |
| L3 (Simple Standing)  | $\frac{d\epsilon}{dt} = -q\epsilon^2 + \ldots$  | Stable                       |
| L4                  | $\frac{d\epsilon}{dt} = 0 + \ldots$            | Neutral                      |
| L5                  | $\frac{d\epsilon}{dt} = q\epsilon + \ldots$      | Unstable                     |
| L6 (Stern Judging)   | $\frac{d\epsilon}{dt} = q\epsilon + \ldots$      | Unstable                     |
| L7 (Staying)         | $\frac{d\epsilon}{dt} = 0 + \ldots$            | Neutral                      |
| L8 (Judging)         | $\frac{d\epsilon}{dt} = q\epsilon + \ldots$      | Unstable                     |
| Image Scoring        | $\frac{d\epsilon}{dt} = 0 + \ldots$            | Neutral                      |

(Simple Standing), and zero for L4, L7, and Image Scoring. The point is that $c_{k_4}$ does not grow over time, so that the one-dimensional dynamics in Eq. (11) remains valid when we consider the leading eight and Image Scoring. If we plug the partial derivatives in Tables 3 and 4 into Eq. (11), the dynamics of $\epsilon$ reduces to

$$\frac{d\epsilon}{dt} \approx \Lambda_1 \epsilon - q(A_{xx} + A_{yz})\epsilon^2.$$ (13)

For Simple Standing, we arrive at

$$\frac{d\epsilon}{dt} \approx -q\epsilon^2,$$ (14)

which admits a solution in the form of $\epsilon \sim 1/t$. Its diverging time scale is self-consistent with our assumption that $\epsilon$ can be approximated as a constant in Eq. (12). As for Image Scoring, on the other hand, $d\epsilon/dt$ is zero up to the second order of perturbation, meaning that the restoring force toward the initial cooperation is still absent. Table 5 shows the results of recovery analysis when applied to each of the leading eight and Image Scoring. These results are consistent with the numerical simulation as shown in the next section.

3. Results

To check the recovery process from error, we conduct numerical simulation. The code is identical to the one used in our previous work [16]: Let us
Figure 1: Recovery from disagreement under the leading eight and Image Scoring (abbreviated as IS). Consistently with Table 5, the average of $\epsilon_{ki}$ converges to zero only under $L_1$ and $L_3$, whereas it remains finite under $L_4$, $L_7$, and IS. The other four norms worsen a small decline in reputation as seen from the gradual increase of $\epsilon_{\text{avg}}$, consistently with $\Lambda_1 > 0$. The black dotted line shows $1/M$ for comparison.
assume that every player uses the same $\alpha$ and $\beta$. We consider a population of size $N \gg 1$ by using an $N \times N$ image matrix, $\{m_{ij}\}$. Initially, the image matrix is filled with ones, and we randomly pick up 20% of the elements and change them to 0.9. At each time step, we pick up a random pair of players $i$ and $j$, the former as the donor and the latter as the recipient. The donor chooses the donation level $\beta_i(m_{ii}, m_{ij})$. This interaction is observed by each of the other members in the society with probability $q$. Each observer, say, $k$, updates $m_{ki}$ according to $\alpha_k[m_{ki}, \beta_i(m_{ii}, m_{ij}), m_{kj}]$. We repeat the above procedure $M$ times, so $M$ can be regarded as a time index.

Our Monte Carlo results in Fig. 1 corroborates the predictions in Table 5. The instability of L2, L5, L6, and L8 is already clear from the fact that $\Lambda_1 > 0$. The stability of L1 and L3 is, however, correctly predicted only when we go through the second-order perturbation theory [Eq. (11)]. The behaviour of $\epsilon \sim 1/t$ is also confirmed by this simulation. Still, the theory leaves the stability of L4 and L7 undetermined, and Fig. 1 suggests that the average of $\epsilon_{ij}$’s will converge to a finite value for them. Image Scoring also belongs to this last category.

4. Discussion and Summary

The continuous dynamics of reputation and behaviour opens up the possibility to apply powerful analytic tools to the study of indirect reciprocity. For this reason, the continuum framework is one of the most convenient ways to study general conditions for norms to be successful when variations in $\alpha$ and $\beta$ can be treated in a perturbative way. A comprehensive and systematic investigation of it would thus greatly enhance our understanding of cooperation through indirect reciprocity. However, our previous linear-order solution provided an inconclusive or incorrect answer when applied to the well-known first-order norm called Image Scoring because the solution did not consider the second-order effects such as justified punishment.

In this work, we have shown how one can go beyond the linear-order analysis. We have expanded the governing equation [Eq. (1)] to the second order of $\epsilon_{ki}$’s and argued the reason that the effective dynamics can reduce to a one-dimensional one. Such reduced dynamics agrees well with numerical simulation for each of the leading eight as well as for Image Scoring. The success of this one-dimensional reduction must be related to the mean-field nature of Eq. (1), although it involves intricate three-body interaction among a donor, a recipient, and an observer at each time step.
Table 6: Characteristics of the leading eight for successful recovery from error. The left two columns show mathematical representations in the continuous and binary models, respectively. The third column means which of the leading eight satisfy the condition, and the last column explains how to interpret the characteristics.

| Continuous | Binary | Norms | Interpretation          |
|------------|--------|-------|-------------------------|
| $\alpha(1, 1, 1) = 1$ | $\alpha_{1C1} = 1$ | L1–L8 | Maintenance of cooperation |
| $\beta(1, 1) = 1$ | $\beta_{11} = C$ | L1–L8 | Forgiveness               |
| $A_x = 0$ | $\alpha_{0C1} = 1$ | L1–L8 | Apology                  |
| $B_x = 0$ | $\beta_{01} = C$ | L1–L8 | Identification of defectors |
| $A_y = 1$ | $\alpha_{1D1} = 0$ | L1–L8 | Punishment                |
| $B_y = 1$ | $\beta_{10} = D$ | L1–L8 | Justification of punishment |
| $A_{yz} > 0$ | $\alpha_{1D0} = 1$ | L1–L8 | Approval for cooperation |
| $A_z = 0$ | $\alpha_{1C0} = 1$ | L1,L3,L4,L7 | to the ill-reputed |
| $A_{zz} = 0$ | $\alpha_{0C0} = 1$ | L1,L3 | to the ill-reputed |

Let us mention a few points on the leading eight from the viewpoint of our second-order perturbation theory (see Table 6). In our continuum framework, all of the leading eight have

$$A_y = B_y = 1$$

$$A_x = B_x = 0$$

and these slope values are related to the basic properties for being nice, retaliatory, apologetic, and forgiving [7]. If we look further into their second derivatives, we find another common feature that they all have

$$A_{yz} > 0$$

so as to justify punishment on an ill-reputed player (Appendix C). The linear-order perturbation theory shows that L2, L5, L6, and L8 are nevertheless unstable because $\Lambda_1 = qA_z > 0$ [see Eq. (8)]. This observation imposes an additional condition that they all have

$$A_z = 0$$

which means that a well-reputed player’s cooperation to an ill-reputed player should be regarded as good. Among the leading eight, L1, L3, L4, and L7 share this property, and we note that they actually show the quickest recovery from error in the original discrete version of the model [15]. Now when it
comes to L4 and L7, which lack restoring force in spite of all the above three conditions [Eqs. (15) to (17)], it turns out that the effect of justified punishment is exactly cancelled out by $A_{xx} = -1$ [see Eq. (13)]. This second derivative is related with how to judge cooperation between two ill-reputed players (see $\alpha_{0C_0}$ in Table 1). L4 and L7 regard such cooperation as bad ($\alpha_{0C_0} = 0$) and fail to recover from error. When it is regarded as good ($\alpha_{0C_0} = 1$), the recovery process succeeds, albeit slowly, as we see from L1 and L3. Our continuum approach thus suggests that the following property stabilizes L1 and L3 in the second-order analysis:

$$A_{xx} = 0.$$  \hspace{1cm} (18)

Together with Eq. (17), this last condition implies that cooperation to an ill-reputed player should be regarded as good, irrespective of the donor’s reputation. All these features are shared by L1 and L3 in common, and their sole difference lies in $\alpha_{0D_0}$ (Table 1). If we focus on L3 (Simple Standing), when you encounter an ill-reputed player, it is always good for your reputation whether you choose to cooperate or punish the co-player. Only defection against a well-reputed player is regarded as bad ($\alpha_{1D_1} = \alpha_{0D_1} = 0$). By disregarding irrelevant information on reputation, Simple Standing achieves a high level of stability in a noisy environment.

One of the open issues that remain untouched in this paper is the evolutionary stability against mutants. Our previous paper analytically obtained the critical benefit-to-cost ratio above which close mutants are driven out, and we may improve the theory by incorporating the second-order effects properly as is done in this paper. This is a promising direction to deepen our understanding of the mechanism to sustain cooperation.

**Acknowledgements**

S.K.B. acknowledges support by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A2071670). Y.M. acknowledges support from Japan Society for the Promotion of Science (JSPS) (JSPS KAKENHI; Grant no. 18H03621 and Grant no. 21K03362). We appreciate the APCTP for its hospitality during the completion of this work.
Appendix A. Second-order perturbation

When \( \epsilon_{ij} \)'s are small parameters, the second-order perturbation for \( \beta \) can be written as follows:

\[
\beta(m_{11}, m_{1j}) = \beta(1 - \epsilon_{11}, 1 - \epsilon_{1j}) \quad \text{(Appendix A.1)}
\]

\[
\approx 1 - B_x \epsilon_{11} - B_y \epsilon_{1j} + \frac{1}{2} B_{xx} \epsilon_{11}^2 + B_{xy} \epsilon_{11} \epsilon_{1j} + \frac{1}{2} B_{yy} \epsilon_{1j}^2 \quad \text{(Appendix A.2)}
\]

\[
\equiv 1 - \kappa. \quad \text{(Appendix A.3)}
\]

Here, we write \( \kappa \equiv \kappa^{(1)} + \kappa^{(2)} \), where \( \kappa^{(1)} \equiv B_x \epsilon_{11} + B_y \epsilon_{1j} \) and \( \kappa^{(2)} \equiv -\left( \frac{1}{2} B_{xx} \epsilon_{11}^2 + B_{xy} \epsilon_{11} \epsilon_{1j} + \frac{1}{2} B_{yy} \epsilon_{1j}^2 \right) \) are first- and second-order corrections, respectively. The second-order perturbation for \( \alpha \) is also straightforward:

\[
\alpha[m_{1i}, \beta_i(m_{ii}, m_{ij}), m_{1j}] \approx \alpha(1 - \epsilon_{1i}, 1 - \kappa, 1 - \epsilon_{1j}) \quad \text{(Appendix A.4)}
\]

\[
\approx 1 - A_x \epsilon_{11} - A_y \kappa - A_z \epsilon_{1j} + \frac{1}{2} A_{xx} \epsilon_{1i}^2 + \frac{1}{2} A_{yy} (\kappa^{(1)})^2 + \frac{1}{2} A_{zz} \epsilon_{1j}^2
\]

\[
+ A_{xy} \epsilon_{11} \kappa^{(1)} + A_{yx} \kappa^{(1)} \epsilon_{1j} + A_{xz} \epsilon_{1i} \epsilon_{1j}. \quad \text{(Appendix A.5)}
\]

Appendix B. Eigenvalues of the system of \( c_{ki} \)’s

For \( N = 2 \), the full eigenvalue structure of Eq. (12) is obtained as follows:

\[
\lambda_1 = q[-1 + A_x + A_z - \{A_{xx} + 2A_{xz} + A_{zz} + (A_{xy} + A_{yz})(B_x + B_y)\} \epsilon]
\]

\[
\lambda_2 = q[-1 + A_x - A_z - \{A_{xx} - A_{zz} + (A_{xy} - A_{yz})(B_x + B_y)\} \epsilon]
\]

\[
\lambda_3 = q[-1 + A_x + A_z + A_y B_x + A_y B_y - \{A_{xx} + 2A_{xz} + A_{zz} + 2(A_{xy} + A_{yz})(B_x + B_y) + A_{yy}(B_x + B_y)^2 + A_y(B_{xx} + 2B_{xy} + B_{yy})\} \epsilon]
\]

\[
\lambda_4 = q[-1 + A_x - A_z + A_y B_x - A_y B_y - \{A_{xx} - A_{zz} + 2A_{xy} B_x - 2A_{yz} B_y + A_{yy}(B_x - B_y)(B_x + B_y) + A_y(B_{xx} - B_{yy})\} \epsilon]. \quad \text{(Appendix B.1)}
\]

If we consider the leading eight, we can readily obtain eigenvalues for \( N = 2, \ldots, 5 \) and generalize the patterns: For example, L1 has the following struc-
where the superscripts on each eigenvalue indicates its multiplicity. For L3, we find a similar result:

\[
\begin{align*}
\lambda_1^{(N^2-2N+1)} &= \frac{1}{N-1} \{- (N-1) + \epsilon \} q \\
\lambda_2^{(N-1)} &= -\frac{1}{N-1} \{ N + (N-4)\epsilon \} q \\
\lambda_3^{(N-1)} &= -(1 + \epsilon)q \\
\lambda_4^{(1)} &= -4\epsilon q,
\end{align*}
\]  

(Appendix B.2)

Finally, the structure becomes even simpler for L4 and L7:

\[
\begin{align*}
\lambda_1^{(N^2-N)} &= -(1 - \epsilon)q \\
\lambda_2^{(N-1)} &= -\frac{N}{N-1} (1 - \epsilon)q \\
\lambda_3^{(1)} &= 0.
\end{align*}
\]  

(Appendix B.4)

In every case, the largest eigenvalue is the last one, which is either negative or zero.

Appendix C. Second derivatives

The second derivative of \( \alpha \) with respect to \( y \) and \( z \) at the fixed point \((x, y, z) = (1, 1, 1)\) can be approximated as

\[
A_{yz} \approx \frac{\alpha(1,1,1) - \alpha(1,1-h,1) - \alpha(1,1-h,1) + \alpha(1,1-h,1-h)}{h^2}
\]  

(Appendix C.1)
with a small parameter $h$. To see the meaning of $A_{yz} > 0$ clearly, let us consider a special case where $\alpha$ has no $z$-dependence when $x = y = 1$. Then, the positivity of $A_{yz}$ is equivalent to

$$\alpha(1, 1-h, 1-h) > \alpha(1, 1-h, 1).$$

(Appendix C.2)

In other words, when one reduces the level of cooperation ($y = 1-h$), it is regarded as good if the co-player has bad reputation ($z = 1-h$). We can generally consider the case with $z$-dependence, and the point is that one earns better reputation by punishing an ill-reputed player than not.

Likewise, we can approximate $A_{xx}$ as

$$A_{xx} \approx \frac{\alpha(1, 1, 1) - \alpha(1-h, 1, 1) - \alpha(1, 1, 1-h) + \alpha(1-h, 1, 1-h)}{h^2}.$$  

(Appendix C.3)

Again, let us assume that $\alpha$ has no $x$-dependence when $y = z = 1$ for convenience of explanation. This assumption is especially relevant to the leading eight because they all have $A_x = 0$. Then, the negativity of $A_{xx}$ means the following:

$$\alpha(1-h, 1, 1-h) < \alpha(1, 1, 1-h).$$

(Appendix C.4)

Note that the right-hand side is effectively the same as $\alpha(1, 1, 1) = 1$ for L1, L3, L4, and L7 because they have $A_z = 0$. The inequality in Eq. (Appendix C.4) implies that cooperation ($y = 1$) between two ill-reputed players ($x = z = 1-h$) is regarded as bad by L4 and L7, for which $A_{xx} = -1$.

References

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

[2] I. M. Silver, A. Shaw, Pint-sized public relations: The development of reputation management, Trends Cogn. Sci. 22 (4) (2018) 277–279.

[3] M. A. Nowak, K. Sigmund, Evolution of indirect reciprocity by image scoring, Nature 393 (6685) (1998) 573.

[4] M. A. Nowak, K. Sigmund, Evolution of indirect reciprocity, Nature 437 (7063) (2005) 1291–1298.
[5] O. Leimar, P. Hammerstein, Evolution of cooperation through indirect reciprocity, Proc. R. Soc. B 268 (1468) (2001) 745–753.

[6] H. Ohtsuki, Y. Iwasa, How should we define goodness? – reputation dynamics in indirect reciprocity, J. Theor. Biol. 231 (1) (2004) 107–120.

[7] H. Ohtsuki, Y. Iwasa, The leading eight: social norms that can maintain cooperation by indirect reciprocity, J. Theor. Biol. 239 (4) (2006) 435–444.

[8] Y. Murase, M. Kim, S. K. Baek, Social norms in indirect reciprocity with ternary reputations, Sci. Rep. 12 (1) (2022) 1–15.

[9] S. Uchida, Effect of private information on indirect reciprocity, Phys. Rev. E 82 (3) (2010) 036111.

[10] S. Uchida, T. Sasaki, Effect of assessment error and private information on stern-judging in indirect reciprocity, Chaos Solitons Fractals 56 (2013) 175–180.

[11] J. Olejarz, W. Ghang, M. Nowak, Indirect reciprocity with optional interactions and private information, Games 6 (4) (2015) 438–457.

[12] I. Okada, T. Sasaki, Y. Nakai, Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring, Sci. Rep. 7 (1) (2017) 1–11.

[13] I. Okada, T. Sasaki, Y. Nakai, A solution for private assessment in indirect reciprocity using solitary observation, J. Theor. Biol. 455 (2018) 7–15.

[14] I. Okada, Two ways to overcome the three social dilemmas of indirect reciprocity, Sci. Rep. 10 (1) (2020) 1–9.

[15] C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, M. A. Nowak, Indirect reciprocity with private, noisy, and incomplete information, Proc. Natl. Acad. Sci. USA 115 (48) (2018) 12241–12246.

[16] S. Lee, Y. Murase, S. K. Baek, Local stability of cooperation in a continuous model of indirect reciprocity, Sci. Rep. 11 (2021) 14225.
[17] M. A. Nowak, Five rules for the evolution of cooperation, Science 314 (5805) (2006) 1560–1563.

[18] R. Sugden, The Economics of Rights, Cooperation and Welfare, Blackwell, Oxford, 1986.

[19] H. Brandt, H. Ohtsuki, Y. Iwasa, K. Sigmund, A survey of indirect reciprocity, in: Y. Takeuchi, Y. Iwasa, K. Sato (Eds.), Mathematics for ecology and environmental sciences, Springer, Berlin, 2007, p. 30.