

# Response of social norms to individual differences in error-proneness

Quang Anh Le<sup>1</sup> and Seung Ki Baek<sup>2,\*</sup>

<sup>1</sup>*Industry-University Cooperation Foundation,  
Pukyong National University, Busan 48513, South Korea*

<sup>2</sup>*Department of Scientific Computing,  
Pukyong National University, Busan 48513, South Korea*

(Dated: March 3, 2025)

## Abstract

Indirect reciprocity explains the evolution of cooperation by considering how our cooperative behavior toward someone is reciprocated by someone else who has observed us. A cohesive society has a shared norm that prescribes how to assess observed behavior as well as how to behave toward others based on the assessments, and the eight social norms that are evolutionarily stable against the invasion of mutants with different behavioral rules are referred to as the leading eight, whose member norms are called L1 to L8, respectively. Among the leading eight, L8 (also known as ‘Judging’) has been deemed mostly irrelevant due to its poor performance in maintaining cooperation when each person may have a different opinion about someone instead of forming a public consensus. In this work, we propose that L8 can nevertheless be best protected from assessment errors among the leading eight if we take into account the intrinsic heterogeneity of error proneness among individuals because this norm heavily punishes those who are prone to errors in following its assessment rule. This finding suggests that individual differences should be explicitly taken into account as quenched randomness to obtain a thorough understanding of a social norm working in a heterogeneous environment.

---

\* seungki@pknu.ac.kr

## I. INTRODUCTION

Indirect reciprocity is one of the main mechanisms for promoting the evolution of cooperation through the interaction of assessment and behavior [1–3]. The conditions for a social norm to maintain cooperation against external perturbations have been revealed since the discovery of the ‘leading eight’ (Table I) [4, 5]. Each norm in the leading eight achieves stable cooperation against the invasion of mutant norms with different behavioral rules, as long as the society has a public consensus in assessing each individual. This line of research has been extended to a more realistic situation in which assessments are made privately, instead of forming a consensus, and it has turned out that privateness renders many of the leading eight untenable [6, 7]. In particular, we have seen that strict norms such as Judging and Stern Judging (also known as L8 and L6, respectively) fail to sustain cooperation when disagreement arises [8, 9]. For example, Judging divides an all-to-all connected society into a weakly balanced configuration [10] with many antagonistic groups [11], resulting in the lowest level of cooperation among the leading eight [12]. However, weakly balanced structures are often found in empirical social networks [13, 14], suggesting a possible selective advantage of Judging. In this work, we wish to explain why Judging is a special norm in terms of robustness in a noisy environment.

From a methodological point of view, the assumption of private assessment can be viewed as an attempt to go beyond a mean-field approach, which has usually been used by considering a well-mixed population (see, however, Ref. [7] demonstrating the importance of population structure in indirect reciprocity). The mean-field approximation greatly reduces the number of degrees of freedom to make the problem tractable, but the price is that it loses every piece of information about individual differences. As a way of retaining individual differences, this work assumes that each individual has a different probability of error. Some are hasty in assessing others, while some others are more prudent and seldom make mistakes in recognizing someone’s goodness. The existence of such individual differences is obvious, and it sounds plausible that erroneous defection will be harmful to one’s own reputation in a cooperative society. However, to our knowledge, the correlation between an individual’s error probability and his or her overall reputation in the long run has not yet been investigated. In this work, we will present some analytic progresses on this issue.

Before proceeding, let us classify errors into three types. The first is an assessment

TABLE I. The leading eight. Each of the eight norms has an assessment rule  $\alpha$  and a behavioral rule  $\beta$ . An observer is observing an interaction between a donor and a recipient, where the donor chooses behavior between cooperation (C) and defection (D), and the observer regards the donor's behavior as either good (G) or bad (B). The assessment rule tells the observer to assign  $\alpha_{XYZ} \in \{G, B\}$  to the donor when the observer regards the donor as  $X \in \{G, B\}$ , the donor does  $Y \in \{C, D\}$  to the recipient, and the observer regards the recipient as  $Z \in \{G, B\}$ . The donor chooses behavior  $\beta_{XY} \in \{C, D\}$  to the recipient when the donor's self-assessment is  $X \in \{G, B\}$  and the donor regards the recipient as  $Y \in \{G, B\}$ .

	$\alpha_{GCG}$	$\alpha_{GDG}$	$\alpha_{GCB}$	$\alpha_{GDB}$	$\alpha_{BCG}$	$\alpha_{BDG}$	$\alpha_{BCB}$	$\alpha_{BDB}$	$\beta_{GG}$	$\beta_{GB}$	$\beta_{BG}$	$\beta_{BB}$
L1	G	B	G	G	G	B	G	B	C	D	C	C
L2 (Consistent Standing)	G	B	B	G	G	B	G	B	C	D	C	C
L3 (Simple Standing)	G	B	G	G	G	B	G	G	C	D	C	D
L4	G	B	G	G	G	B	B	G	C	D	C	D
L5	G	B	B	G	G	B	G	G	C	D	C	D
L6 (Stern Judging)	G	B	B	G	G	B	B	G	C	D	C	D
L7 (Staying)	G	B	G	G	G	B	B	B	C	D	C	D
L8 (Judging)	<i>G</i>	<i>B</i>	<i>B</i>	<i>G</i>	<i>G</i>	<i>B</i>	<i>B</i>	<i>B</i>	C	D	C	D

error, which means that an observer remembers a donor as good, although the correct assessment should be the opposite, or vice versa. This is the type of error that we will focus on throughout this work. The second is a perception error, by which an observer mistakes a donor's cooperation as defection, or vice versa. Note the difference between the assessment error and the perception error: If the observer is an unconditional cooperator, the perception error does not change the observer's assessment, whereas the assessment error does. However, if we work with the leading eight in the vicinity of paradise where everyone is good, the assessment is heavily based on the observed behavior, so the perception error plays a similar role to that of the assessment error. The last is a behavioral error. A donor cooperates by error, although the correct behavior is defection, or vice versa. Later, we will examine the effects of behavioral errors through numerical calculations.

## II. ANALYSIS

We consider a population of size  $N \gg 1$ . Let  $m_{ij}^t$  denote how an individual  $i$  assesses an individual  $j$  at time  $t$ . If  $i$  regards  $j$  as perfectly good (bad), we have  $m_{ij}^t = 1(0)$ , but it is generally between zero and one [8, 15–17]. The conventional discrete model can be said to use only the end points. The system becomes more analytically tractable when we work with continuous variables. Even if we consider the discrete model, we expect that the continuous description can capture the average behavior involved with probabilistic errors. To see the correspondence between these two approaches, we will carry out analytic calculations within the continuous model, while the discrete model is used in numerical simulations.

For each round, a pair of randomly chosen individuals  $i$  and  $j$  interact with each other by playing the donation game, where  $i$  as a donor can choose to cooperate or defect. If the donor cooperates, the donor's payoff decreases by  $c$  as the cost of cooperation, and the other individual  $j$  playing the role of the recipient earns  $b$  as the benefit of cooperation. However, if the donor defects, it means that the donor refuses to cooperate and their payoffs do not change. When  $b > c > 0$ , the donation game is a special type of prisoner's dilemma. What  $i$  does to  $j$  is determined by his or her behavior rule  $\beta_i$ , which depends on  $i$ 's self-assessment as well as on how  $i$  regards  $j$ . Mathematically speaking, this can be expressed by  $\beta_i = \beta_i(m_{ii}^t, m_{ij}^t)$ . In the continuous version, the donor  $i$  pays the cost of cooperation  $c\beta_i$ , which benefits the recipient by  $b\beta_i$ , so that  $\beta_i = 1$  and  $0$  mean full cooperation and defection, respectively. Every observer  $k$  observes the interaction between  $i$  and  $j$  with probability  $q$  and assesses the donor  $i$  according to his or her own assessment rule  $\alpha_k$ . The assessment depends on how  $k$  regards  $i$ , what  $i$  does to  $j$ , and how  $k$  regards  $j$ , which can be expressed by  $\alpha_k = \alpha_k[m_{ki}^t, \beta_i(m_{ii}^t, m_{ij}^t), m_{kj}^t]$ . However, with probability  $\sigma_{ki}$ , the assessment can be flipped to  $1 - \alpha_k$ . An individual  $k$ 's social norm is the combination of the assessment rule  $\alpha_k$  and the behavioral rule  $\beta_k$ . Table II shows the continuous versions of the leading eight obtained through bilinear and trilinear interpolations so that the original definitions are recovered at the end points.

The above dynamical rule can be written as the following equation:

$$m_{ki}^{t+1} = (1 - q)m_{ki}^t + \frac{q}{N} \sum_{j=1}^N (1 - \sigma_{ki}) \alpha_k [m_{ki}^t, \beta_i(m_{ii}^t, m_{ij}^t), m_{kj}^t] + \sigma_{ki} \{1 - \alpha_k [m_{ki}^t, \beta_i(m_{ii}^t, m_{ij}^t), m_{kj}^t]\}, \quad (1)$$

TABLE II. Continuous expressions of the leading eight obtained through bi- and tri-linear interpolations. The last column shows the value of  $\alpha$  when  $m_{ki} = 1/2$  for every pair of  $k$  and  $i$ .

Norm	$\alpha(x, y, z)$	$\beta(x, y)$	$\alpha^* \equiv \alpha \left[ \frac{1}{2}, \beta \left( \frac{1}{2}, \frac{1}{2} \right), \frac{1}{2} \right]$
L1	$x + y - xy - xz + xyz$	$-x + xy + 1$	13/16
L2 (Consistent Standing)	$x + y - 2xy - xz + 2xyz$	$-x + xy + 1$	5/8
L3 (Simple Standing)	$yz - z + 1$	$y$	3/4
L4	$-y - z + xy + 2yz - xyz + 1$	$y$	5/8
L5	$-z - xy + yz + xyz + 1$	$y$	5/8
L6 (Stern Judging)	$-y - z + 2yz + 1$	$y$	1/2
L7 (Staying)	$x - xz + yz$	$y$	1/2
L8 (Judging)	$x - xy - xz + yz + xyz$	$y$	3/8

where  $\sigma_{ki}$  is the probability of assessment error, which may depend on the observer  $k$  or the donor  $i$ . In the second term on the right-hand side, we have taken the average over the randomly chosen recipient  $j$ , which may also be equal to  $i$  for mathematical convenience. Assuming that everyone uses the same norm, we can say  $\alpha_k = \alpha$  and  $\beta_i = \beta$  without the subscripts. Rearranging the terms of Eq. (1) in the long-time limit, we have the following  $N^2$ -dimensional system of equations to solve:

$$0 = -m_{ki} + \sigma_{ki} + \frac{(1 - 2\sigma_{ki})}{N} \sum_{j=1}^N \alpha [m_{ki}, \beta(m_{ii}, m_{ij}), m_{kj}], \quad (2)$$

where  $\alpha$  and  $\beta$  are given in Table II and the superscripts can now be neglected. Note that the observation probability  $q$  becomes irrelevant in this steady state. It is clearly seen that  $m_{ki} = 1/2$  if  $\sigma_{ki} = 1/2$ , which means that the observer makes random assessments.

### A. How an error-prone individual assesses others

As a specific example, consider L3. We will furthermore assume that the error probability depends only on the observer so that  $\sigma_{ki} = \sigma_k$ . Equation (2) is then rewritten as

$$0 = -m_{ki} + \sigma_k + (1 - 2\sigma_k) (C_{ki} - \mu_k + 1), \quad (3)$$

where  $C_{ki} \equiv N^{-1} \sum_j m_{kj} m_{ij}$  and  $\mu_k \equiv N^{-1} \sum_j m_{kj}$ . By assuming that  $(m_{kj} - \mu_k)$  and  $(m_{ij} - \mu_i)$  fluctuate independently, we replace  $C_{ki}$  by  $\mu_k \mu_i$  to get

$$0 \approx -m_{ki} + \sigma_k + (1 - 2\sigma_k)(\mu_k \mu_i - \mu_k + 1). \quad (4)$$

Summing both sides of Eq. (3) over  $i$  and dividing them by  $N$ , we obtain

$$0 \approx -\mu_k + \sigma_k + (1 - 2\sigma_k)(\mu_k \bar{\mu} - \mu_k + 1), \quad (5)$$

where  $\bar{\mu} \equiv N^{-1} \sum_i \mu_i$ . We postulate that how an observer  $k$  assesses others is determined by the observer's probability of error, so that we can write  $\mu_k = \mu(\sigma_k)$ . The simplest functional form would be a linear function such as  $\mu(\sigma_k) = u\sigma_k + v$  with constants  $u$  and  $v$ . To satisfy  $\mu(\sigma_k = 1/2) = 1/2$ , we have  $v = (1 - u)/2$ , which means that

$$\mu_k = \mu(\sigma_k) = u\sigma_k + \frac{1}{2}(1 - u). \quad (6)$$

If  $\sigma_k$  is uniformly distributed between 0 and 1/2, we should have

$$\bar{\mu} = \frac{1}{4}(2 - u) \quad (7)$$

in the large- $N$  limit because

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N \sigma_k^n = \frac{\int_0^{1/2} x^n dx}{\int_0^{1/2} dx} = \frac{1}{(n+1)2^n}. \quad (8)$$

Substituting Eqs. (6) and (7) into Eq. (5) and averaging both the sides over  $k$  by using Eq. (8), we get an algebraic equation for  $u$ , whose physical solution is  $u = -1/2$ . Thus, Eq. (6) in this case predicts

$$\mu(\sigma_k) = \frac{1}{4}(3 - 2\sigma_k). \quad (9)$$

When  $k = i$ , Eq. (4) leads to

$$m_{kk} \approx \sigma_k + (1 - 2\sigma_k)(\mu_k^2 - \mu_k + 1), \quad (10)$$

by assuming that  $C_{kk} - \mu_k^2$ , i.e., the variance of  $m_{kk}$ , vanishes. Even if this assumption cannot be justified in general, suppose that it is valid in the vicinity of  $\sigma_k = 1/2$ , where  $m_{ki}$  is identically equal to 1/2. Regarding  $m_{kk}$  as a proxy of  $\mu_k$ , we can explicitly solve Eq. (10) for  $\mu_k$  and obtain

$$\mu_k = \frac{1 - \sigma_k - \sqrt{\sigma_k(1 - \sigma_k)}}{1 - 2\sigma_k} = \frac{1}{4}(3 - 2\sigma_k) + O\left(\left|\frac{1}{2} - \sigma_k\right|^3\right), \quad (11)$$

TABLE III. First-order derivatives of  $\alpha$  and  $\beta$  when  $m_{ki} = 1/2$  for every pair of  $k$  and  $i$ .

Norm	$A'_x$	$A'_y$	$A'_z$	$B'_x$	$B'_y$
L1	1/8	3/4	-1/8	-1/2	1/2
L2 (Consistent Standing)	-1/4	1/2	1/4	-1/2	1/2
L3 (Simple Standing)	0	1/2	-1/2	0	1
L4	1/4	1/4	-1/4	0	1
L5	-1/4	1/4	-1/4	0	1
L6 (Stern Judging)	0	0	0	0	1
L7 (Staying)	1/2	1/2	0	0	1
L8 (Judging)	1/4	1/4	1/4	0	1

reproducing Eq. (9) near  $\sigma_k = 1/2$ . We stress that replacing  $m_{kk}$  by  $\mu_k$  is only an approximation to obtain  $\mu_k$ , which is actually dominated by  $m_{ki}$  with  $i \neq k$ .

A convenient way to solve such a nonlinear equation is the Newton method [18]. To see a one-dimensional example for solving  $f(x) = 0$ , let us denote a trial solution as  $\hat{x}$ , while the unknown true solution is denoted as  $x^*$ . We expand the equation around the trial solution to the first order as follows:

$$0 = f(x^*) = f(\hat{x}) + (x^* - \hat{x}) \left. \frac{df}{dx} \right|_{\hat{x}} + \dots, \quad (12)$$

and we observe that the true solution is approximated as

$$x^* \approx \hat{x} - \frac{1}{(df/dx)|_{\hat{x}}} f(\hat{x}). \quad (13)$$

In our problem, the trial solution should be  $\hat{\mu}_k = 1/2$ , which is an exact solution for  $\sigma_k = 1/2$ . To apply the Newton method, we need first-order derivatives evaluated at this trial solution. Let us rewrite  $m_{kk} \approx \mu_k = 1/2 + \delta_k$  and expand Eq. (10) to the first order of  $\delta_k$  as follows:

$$0 = - \left( \frac{1}{2} + \delta_k \right) + \sigma_k + (1 - 2\sigma_k) \alpha \left[ \frac{1}{2} + \delta_k, \beta \left( \frac{1}{2} + \delta_k, \frac{1}{2} + \delta_k \right), \frac{1}{2} + \delta_k \right] \quad (14)$$

$$\approx - \left( \frac{1}{2} + \delta_k \right) + \sigma_k + (1 - 2\sigma_k) \left[ \alpha^* + A'_x \delta_k + A'_y (B'_x \delta_k + B'_y \delta_k) + A'_z \delta_k \right], \quad (15)$$

where  $\alpha^* \equiv \alpha \left[ \frac{1}{2}, \beta \left( \frac{1}{2}, \frac{1}{2} \right), \frac{1}{2} \right]$ ,  $A'_\xi \equiv (\partial\alpha/\partial\xi)_{(x,y,z)=(\frac{1}{2},\beta(\frac{1}{2},\frac{1}{2}),\frac{1}{2})}$ , and  $B'_\xi \equiv (\partial\beta/\partial\xi)_{(x,y)=(\frac{1}{2},\frac{1}{2})}$  (see Table III). Equation (13) then yields

$$\mu_k^* \approx \frac{1}{4} (3 - 2\sigma_k), \quad (16)$$

TABLE IV. Ranges of  $m_{ki}$  that an observer  $i$  receives from others under different social norms, when the probability of error depends on the observer, i.e.,  $\sigma_{ij} = \sigma_i$ . The ranges are obtained by applying the Newton method [20].

	$N = 2$	$N = 3$
L1	$\frac{1}{2} \leq m_{ki} \leq \frac{13}{16}$	$\frac{1}{2} \leq m_{ki} \leq \frac{13}{16}$
L2 (Consistent Standing)	$\frac{1}{2} \leq m_{ki} \leq \frac{5}{8}$	$\frac{1}{2} \leq m_{ki} \leq \frac{5}{8}$
L3 (Simple Standing)	$\frac{1}{2} \leq m_{ki} \leq \frac{27-15\sigma_i}{36-12\sigma_i}$	$\frac{1}{2} \leq m_{ki} \leq \frac{27-18\sigma_i}{36-16\sigma_i}$
L4	$\frac{1}{2} \leq m_{ki} \leq \frac{18+3\sigma_i}{27+9\sigma_i}$	$\frac{1}{2} \leq m_{ki} \leq \frac{18+\sigma_i}{27+6\sigma_i}$
L5	$\frac{1}{2} \leq m_{ki} \leq \frac{135-78\sigma_i}{225-120\sigma_i}$	$\frac{1}{2} \leq m_{ki} \leq \frac{135-84\sigma_i}{225-130\sigma_i}$
L6 (Stern Judging)	$m_{ki} = \frac{1}{2}$	$m_{ki} = \frac{1}{2}$
L7 (Staying)	$m_{ki} = \frac{1}{2}$	$m_{ki} = \frac{1}{2}$
L8 (Judging)	$\frac{1}{2} \leq m_{ki} \leq \frac{12\sigma_i}{9+36\sigma_i}$	$\frac{1}{2} \leq m_{ki} \leq \frac{10\sigma_i}{9+30\sigma_i}$

in agreement with Eq. (9). As for the other norms, the general expression is given as follows:

$$\mu_k^* \approx \frac{1}{2} \left\{ 1 + \frac{(2\alpha^* - 1)(1 - 2\sigma_k)}{1 - (1 - 2\sigma_k) [A'_x + A'_y(B'_x + B'_y) + A'_z]} \right\}. \quad (17)$$

This formula predicts that only L8 will exhibit positive correlation between  $\sigma_k$  and  $\mu_k$ , and the reason is that only L8 has  $\alpha^* < 1/2$  (Table II). The correlation vanishes for L6 and L7 because  $\alpha^* = 1/2$ . The behavior of L6 is totally driven by entropy, as has already been analyzed in detail [9]. Concerning L7, as long as  $\alpha^* = 1/2$ , the steady-state equation [Eq. (2)] actually admits a solution such that  $m_{ki} = 1/2$  for every pair of  $k$  and  $i$ , regardless of the distribution of  $\{\sigma_k\}$ . For the other five norms from L1 to L5, the correlation is negative, which means that a careless individual tends to assign low assessments to others. Figure 1 shows that all these predictions are well corroborated by numerical simulations.

## B. How an error-prone individual is assessed by others

As we have already seen, when assessment errors are caused by observers, that is,  $\sigma_{ki} = \sigma_k$ , the dominant factor determining  $m_{ki}$  should be the probability of error of the individual who makes the assessment, i.e.,  $\sigma_k$ , but the working hypothesis here is that  $\sigma_i$  of the one being assessed can also affect the assessment in the long run. Our analysis given above effectively



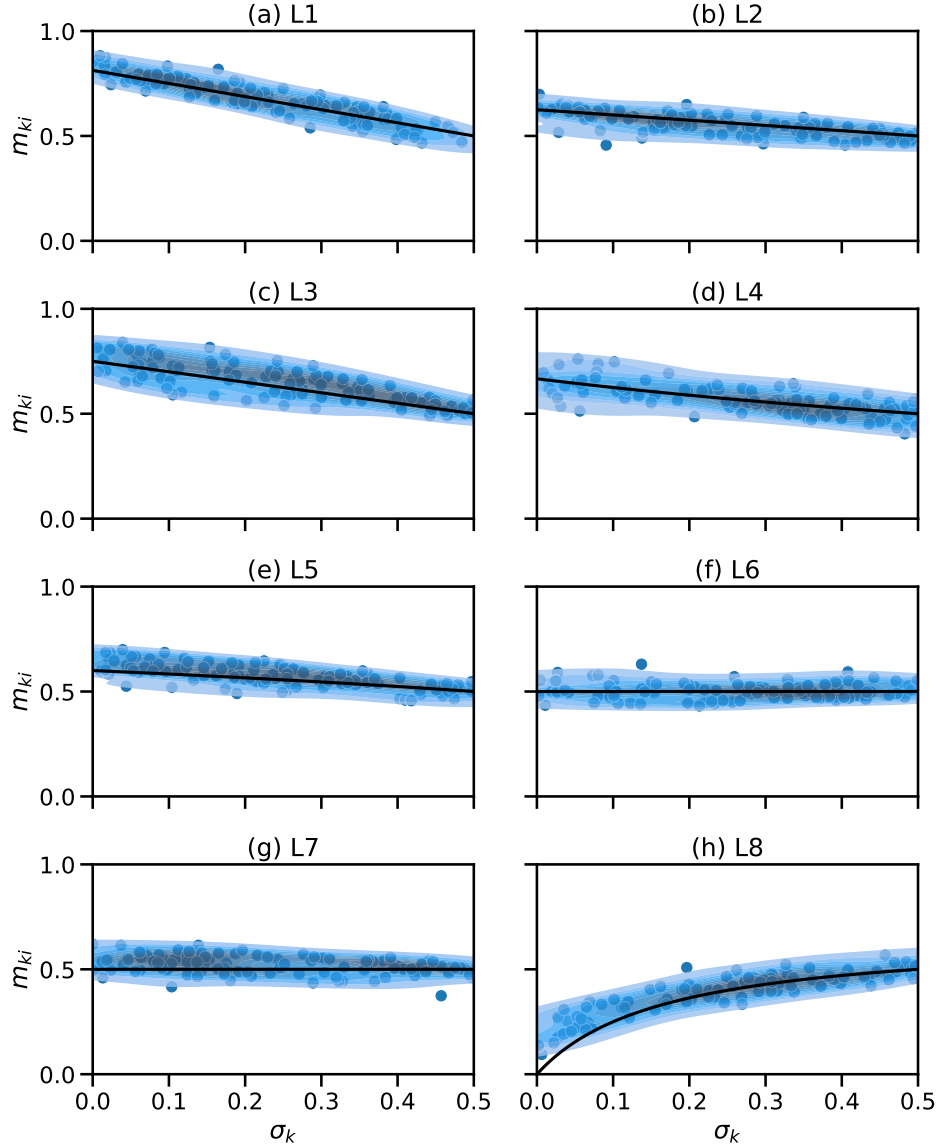


FIG. 1. Numerical results from the discrete model in which  $m_{ki}$  can take only 0 and 1. The size of the population is  $N = 10^2$ , and the observation probability is  $q = 1$ . The horizontal axis is the probability of assessment error of an individual  $k$ , and the vertical axis is the time average of  $m_{ki}$  for  $T = 5 \times 10^4$  rounds, after discarding transient data for the first  $T$  rounds. In every panel, we have  $10^2$  data points, each of which has been obtained by picking up two specific individuals  $k$  and  $i$  from a different sample. Each sample runs with an independent realization of  $\{\sigma_k\}$  as a set of quenched random numbers between 0 and  $1/2$ . As the initial condition, we fill the image matrix  $\{m_{ki}\}$  with random numbers uniformly distributed in the unit interval. The shades show the kernel-density estimates [19], and the solid lines are obtained from Eq. (17).

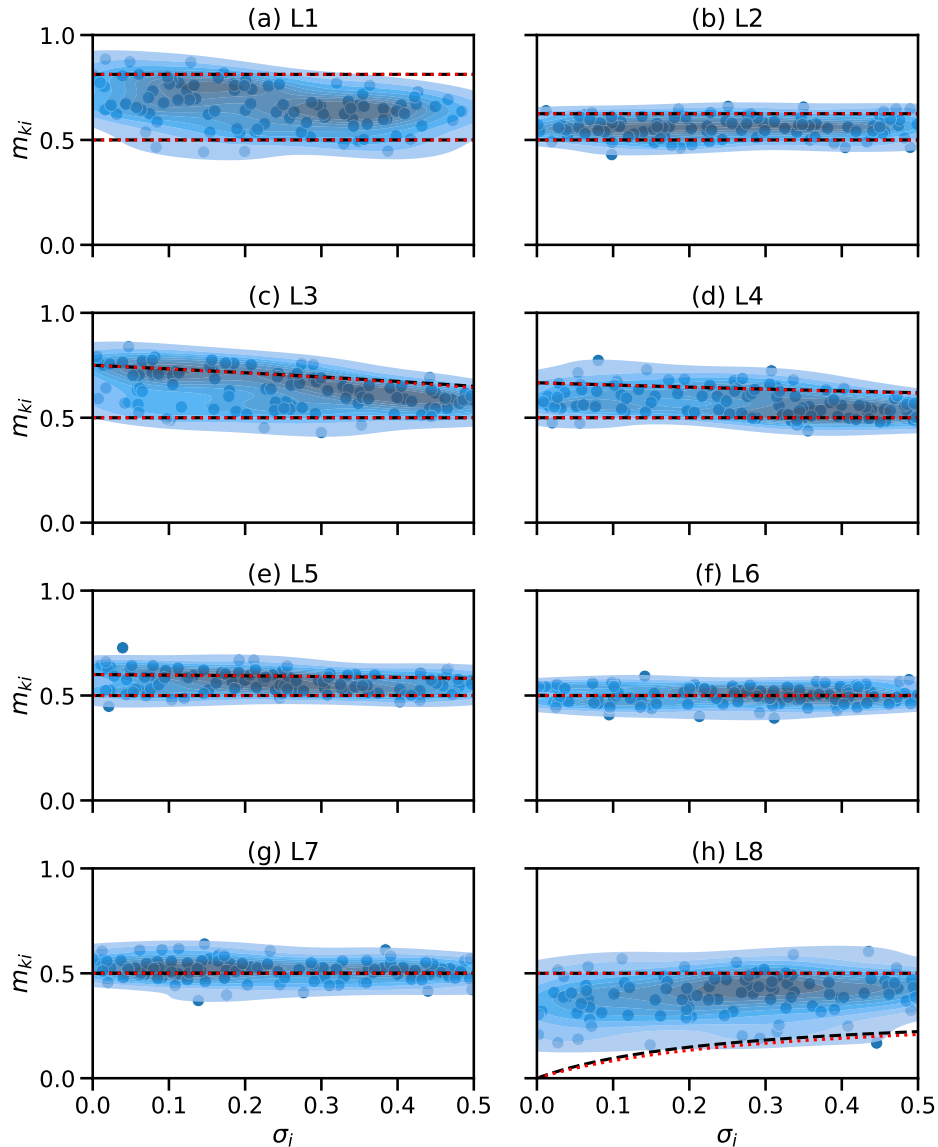


FIG. 2.  $m_{ki}$  plotted against  $\sigma_i$ , instead of  $\sigma_k$ . We have used the same data as in Fig. 1. The black dashed lines show the cases of  $N = 2$ , and the red dotted lines show the cases of  $N = 3$  in Table IV.

corresponds to a one-body problem [see, e.g., the derivation of Eq. (11)], which can be obtained from Eq. (2) by taking  $N = 1$ . To estimate how others assess an individual with a finite error probability, we need  $N = 2$  at least, and we will apply the Newton method again to find an approximate solution. In a  $d$ -dimensional problem given by  $f_1(x_1, x_2, \dots, x_d) =$

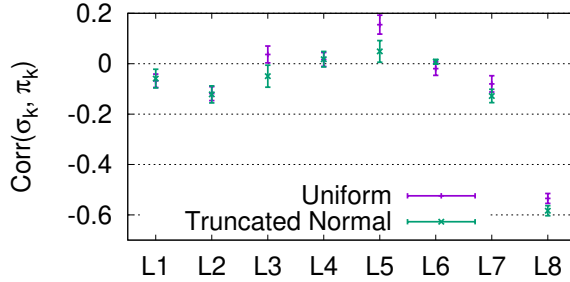


FIG. 3. Pearson correlation between an individual  $k$ 's payoff  $\pi_k$  and the error probability  $\sigma_k$ , when the benefit of cooperation is  $b = 1$  whereas the cost is  $c = 1/2$ . The simulation details are the same as in Fig. 1. We have recorded a specific individual's payoff [Eq. (19)] at the end of each run and calculated the correlation coefficient from  $10^2$  such data points. To estimate the mean and the standard error, we have repeated this procedure 10 times. The error probabilities  $\{\sigma_k\}$  are drawn from  $[0, \frac{1}{2})$  either uniformly or by following the truncated normal distribution whose center and width are 0 and  $1/4$ , respectively.

$f_2(x_1, x_2, \dots, x_d) = \dots = f_d(x_1, x_2, \dots, x_d) = 0$ , Eq. (13) is rewritten as

$$\begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_d^* \end{pmatrix} \approx \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_d \end{pmatrix} - \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_d} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_d}{\partial x_1} & \frac{\partial f_d}{\partial x_2} & \dots & \frac{\partial f_d}{\partial x_d} \end{pmatrix}^{-1} \begin{pmatrix} f_1(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_d) \\ f_2(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_d) \\ \vdots \\ f_d(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_d) \end{pmatrix}, \quad (18)$$

where the terms on the right-hand side are all evaluated at the trial solution  $(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_d)$ . The resulting solution  $\{m_{ki} = m_{ki}^*\}$  generally depends on both  $\sigma_k$  and  $\sigma_i$ , and we present the range of  $m_{ki}$  for each norm as a function of  $\sigma_i$  to remove the dependency on  $\sigma_k$  in Table IV. We have performed the analytic calculation only for  $N = 2$  and  $3$ , but the size dependence appears to be weak, and the predicted ranges are qualitatively consistent with the numerical data (Fig. 2).

### III. DISCUSSION

The important point of the above analysis is that the correlation of  $m_{ki}$  with  $\sigma_i$  is found to be weaker than the correlation with  $\sigma_k$ . For example, if we look at L8, the amount of donation that an individual  $k$  gives to others increases with  $\sigma_k$  [Fig. 1(h)], while the amount

of cooperation that he or she receives is almost insensitive to  $\sigma_k$  [Fig. 2(h)]. If we define an individual's payoff in the following way,

$$\pi_k \equiv \frac{1}{N} \sum_{i=1}^N [b\beta (m_{ii}, m_{ik}) - c\beta (m_{kk}, m_{ki})], \quad (19)$$

it is thus expected to decrease under the action of L8 as  $\sigma_k$  grows. This prediction is verified by our numerical results (Fig. 3). Among the leading eight, L8 exhibits the clearest signal of negative correlation between  $\sigma_k$  and  $\pi_k$ . The more often one makes mistakes, the lower the payoff. The correct assessment is that others are bad [Fig. 1(h)], so the level of cooperation is low. In contrast, even if L3 shows a higher level of cooperation [8, 12], Fig. 3 shows that it does not impose a heavy penalty for error-proneness. According to Eq. (17), the decisive factor for determining the sign of the correlation is  $\alpha^* \equiv \alpha \left[ \frac{1}{2}, \beta \left( \frac{1}{2}, \frac{1}{2} \right), \frac{1}{2} \right]$ , which makes sense in the continuous model where everyone can have a half-good state, but is also proved useful in the discrete model (Fig. 1), showing the power of the continuous model. If the social norm cannot support the half-good state by itself, that is, if  $\alpha^* < 1/2$ , a careful individual with low  $\sigma_k$  will regard others as bad, securing his or her own payoff, although the average level of cooperation remains finite in the population. Although the derivation in Eq. (8) uses the information on how the error probabilities are distributed in the population, Fig. 3 shows that L8 still has the most negative correlation even if the distribution is not uniform. We also observe the same trend for larger populations (not shown).

As for behavioral errors, by which one chooses defection although cooperation is intended and vice versa, we may expect that all norms in the leading eight will punish error-prone individuals because they have been designed to suppress behavioral mutants [3], which can mimic behavioral errors. Our numerical calculations show that this is indeed the case, with two trivial exceptions having negligible correlations. One is L6, totally driven by entropy even with an arbitrarily small error probability [9], and the other is L8, under which everyone is eventually considered bad.

#### IV. SUMMARY

In summary, we have investigated how a social norm shapes society in the presence of heterogeneity in individual probabilities of assessment errors. So far, the stability of a social norm has usually been analyzed in terms of error and mutation, but one of the

theoretical challenges is to deal with these sources of randomness within a reasonable number of degrees of freedom. Concerning error, the traditional approach assumes that everyone is equally prone to it. When it comes to mutation, although it introduces heterogeneity in the population, most studies restrict themselves to the low mutation limit to simplify the problem (see, however, Ref. [21] as an exception).

This work has extended the concept of stability by considering individual heterogeneity in assessment errors as a quenched disorder and has shown the possibility of analytic understanding. Among the leading eight, L8 (Judging) strongly punishes those who do not carefully follow the norm, and it suggests the unique power of Judging when there exists heterogeneity among individuals. The important factor turns out to be how the norm assesses the situation where everyone is half good, although it is a hypothetical point that makes full sense when the norm is interpolated between good and bad. It confirms the necessity and usefulness of the continuous model of indirect reciprocity.

## ACKNOWLEDGMENTS

We acknowledge support by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A2071670).

- 
- [1] M. A. Nowak and K. Sigmund, *Nature* **393**, 573 (1998).
  - [2] M. A. Nowak and K. Sigmund, *Nature* **437**, 1291 (2005).
  - [3] H. Ohtsuki and Y. Iwasa, *J. Theor. Biol.* **231**, 107 (2004).
  - [4] H. Ohtsuki and Y. Iwasa, *J. Theor. Biol.* **239**, 435 (2006).
  - [5] H. Ohtsuki, Y. Iwasa, and M. A. Nowak, *Nature* **457**, 79 (2009).
  - [6] C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, and M. A. Nowak, *Proc. Natl. Acad. Sci. USA* **115**, 12241 (2018).
  - [7] Y. Murase and C. Hilbe, *Proc. Natl. Acad. Sci. USA* **121**, e2406885121 (2024).
  - [8] S. Lee, Y. Murase, and S. K. Baek, *J. Theor. Biol.* **548**, 111202 (2022).
  - [9] M. Bae, T. Shimada, and S. K. Baek, *Phys. Rev. E* **110**, L052301 (2024).
  - [10] D. Easley and J. Kleinberg, “A weaker form of structural balance,” in *Networks, Crowds,*

- and Markets: Reasoning about a Highly Connected World* (Cambridge University Press, Cambridge, 2010) Chap. 5, pp. 115–118.
- [11] M. Bae and S. K. Baek, “Indirect reciprocity as a dynamics for weak balance,” arXiv:2501.05824 (2025).
  - [12] Y. Fujimoto and H. Ohtsuki, *PRX Life* **2**, 023009 (2024).
  - [13] J. Leskovec, D. Huttenlocher, and J. Kleinberg, in *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.* (Association for Computing Machinery, New York, NY, 2010) pp. 1361–1370.
  - [14] M. Szell, R. Lambiotte, and S. Thurner, *Proc. Natl. Acad. Sci. USA* **107**, 13636 (2010).
  - [15] S. Lee, Y. Murase, and S. K. Baek, *Sci. Rep.* **11**, 14225 (2021).
  - [16] Y. Mun and S. K. Baek, *Eur. Phys. J. Spec. Top.* **233**, 1251 (2024).
  - [17] Y. Mun, Q. A. Le, and S. K. Baek, *J. Korean Phys. Soc.* **85**, 969 (2024).
  - [18] S. K. Baek, S. Do Yi, and H.-C. Jeong, *J. Theor. Biol.* **430**, 215 (2017).
  - [19] M. L. Waskom, *J. Open Source Softw.* **6**, 3021 (2021).
  - [20] “Mathematica, Version 10.0,” (Wolfram Research, Inc., Champaign, IL, 2014).
  - [21] V. V. Vasconcelos, F. P. Santos, F. C. Santos, and J. M. Pacheco, *Phys. Rev. Lett.* **118**, 058301 (2017).