On the stability of cooperation under indirect reciprocity with first-order information

Berger, Ulrich; Grüne, Ansgar

Published in:
Games and Economic Behavior

DOI:
10.1016/j.geb.2016.05.003

Published: 01/01/2016

Document Version
Peer reviewed version

On the stability of cooperation under indirect reciprocity with first-order information*

Ulrich Berger† Ansgar Grüne‡

May 22, 2016

Abstract: Indirect reciprocity describes a class of reputation-based mechanisms which may explain the prevalence of cooperation in large groups where partners meet only once. The first model for which this has been demonstrated was the image scoring mechanism. But analytical work on the simplest possible case, the binary scoring model, has shown that even small errors in implementation destabilize any cooperative regime. It has thus been claimed that for indirect reciprocity to stabilize cooperation, assessments of reputation must be based on higher-order information. Is indirect reciprocity relying on first-order information doomed to fail? We use a simple analytical model of image scoring to show that this need not be the case. Indeed, in the general image scoring model the introduction of implementation errors has just the opposite effect as in the binary scoring model: it may *stabilize* instead of destabilize cooperation.

Key words: cooperation; prisoner's dilemma; donation game; indirect reciprocity; image scoring; first-order assessment; first-order information; evolutionary stability;

JEL classification: C72, D83

*Financial support by the Jubiläumsstiftung der Wirtschaftsuniversität Wien is gratefully acknowledged.

†WU Vienna, Department of Economics, Welthandelsplatz 1, 1020 Wien, Austria, ulrich.berger@wu.ac.at

‡Beethovenstr. 55, 53115 Bonn, Germany, ansgar.gruene@gmail.com

1 Introduction

1.1 Indirect reciprocity

Cooperating by acting altruistically and helping others reduces the actor's material payoff and increases the recipient's material payoff. If the sum of the payoffs increases, cooperation enhances welfare and is socially beneficial. But actions which reduce own payoff are hard to reconcile with individual rationality, so why do we see so much cooperation in economic life? Questions such as this one have traditionally been studied using the framework of the Prisoner's Dilemma game, often in the special case of the donation game. In these games defection is the inevitable outcome unless cooperation can be induced by some supporting mechanism. Such mechanisms solve the paradox of cooperation by placing the Prisoner's Dilemma into an environment where interactions occur repeatedly and short-run altruism is rewarded in the long run and can thus become established in a society. Nowak (2006) surveys the most important such mechanisms from the biologist's point of view. For economists, mechanisms based on *reciprocity* are of primary interest. If pairs of players from a large population interact only once, cooperation based on personal enforcement, also called direct reciprocity (Axelrod and Hamilton, 1981, Fudenberg and Maskin, 1990), is ruled out. But cooperation can still be achieved via *indirect reciprocity* (Trivers, 1985, Sugden, 1986, Alexander, 1987). Under indirect reciprocity, helping others enhances one's reputation, and help is primarily directed towards those with a high reputation. Thus, the immediate costs of helping are more than offset by the future benefits of being helped when in need, which aligns individual and social rationality of cooperation.

Economists have studied mechanisms of indirect reciprocity within the framework of repeated games with random matching, where these mechanisms are known as community enforcement.[1] The appropriateness of the traditional approach via repeated games has sometimes been questioned because it relies on perfectly rational players with common knowledge of rationality and perfect anticipation of other's reactions to hypothetical deviations from equilibrium play. While such stringent assumptions might arguably be approximately justified in dyadic long-run interactions, they are considerably less convincing as a model of repeated interactions in large real-world societies with constant entry and exit. Some researchers have therefore turned to an approach relying on evolutionary games (Weibull, 1997, Samuelson, 1998, Gintis, 2000, Cressman 2003, Sandholm, 2010). Borrowed from biology (Maynard Smith, 1982, Hofbauer and Sigmund, 1998), evolutionary game theory posits boundedly rational myopic agents choosing their strategies from a restricted set of rules-of-thumb and occasionally switching to strategies promising higher payoffs. In such a setting, indirect reciprocity has been formally studied since the late 1990ies.[2] The pioneering work of Nowak and Sigmund (1998a, 1998b)

[1] This literature started with Kandori (1992) and was further developed by Ellison (1994), Okuno-Fujiwara and Postlewaite (1995), Takahashi (2010), and Awaya (2014), among others.

[2] While indirect reciprocity is highly relevant for studies of human cooperation, and therefore for eco-

showed that in a population of discriminators who base their decisions on their partner's reputation, cooperation can persist for long periods of time. In their models, reputation is measured as an *image score*.

1.2 Image scoring

Under image scoring, every individual carries an observable numerical score measuring its past cooperativeness by counting how often it helped on its past interactions. Nowak and Sigmund (1998a) studied the performance of discriminatory strategists who cooperate if and only if their opponent's image score exceeds a given threshold. These *threshold strategies* differ in their respective threshold levels and include unconditional defection and unconditional cooperation as extreme cases. If only a single past action of an individual is observed, the score becomes binary and the only proper discriminator strategy assesses other individuals as either *Good* or *Bad*, depending on whether or not they gave help. In any interaction, discriminators then help those and only those which are assessed as Good. In assessing an individual's reputation, the scoring rule relies only on the individual's behavior towards its interaction partner, but neither on this partner's reputation nor on the individual's previous reputation. Such an assessment rule is called a first-order assessment rule.

Image scoring seemed to work well in Nowak and Sigmund's (1998a) numerical simulations in the sense that cooperative regimes could persist for many generations, but cooperation was not evolutionarily stable. Unconditional cooperators could invade by neutral drift and pave the way for defectors to take over the population for some time, leading to endless cycles. The analytical results on four different model specifications of the binary version of image scoring in Nowak and Sigmund (1998b) made it clear that discriminators are only neutrally stable and can always be invaded by unconditional cooperators drifting into the population. All these models assumed that implementation of strategies is free of errors. In principle, introducing noise has the potential to stabilize cooperation, as was pointed out by Boyd (1989) for the case of direct reciprocity. However, for image scoring things seemed to become even worse under errors. Indeed, Panchanathan and Boyd (2003) pointed out that if errors in the implementation of strategies are added to the binary scoring model, cooperation becomes unstable and defection prevails in the long run. The reason for this is the paradoxical nature of image scoring: a discriminator who refuses to help a "bad" opponent risks being assessed as "bad" itself.

nomics, it has been somewhat neglected by economists. With the notable exception of a few experimental economics studies (see the references in section 1.3) the field is dominated by theoretical biologists and anthropologists as well as mathematicians and physicists. It is the authors' hope to make a modest contribution to introducing this area of studies to economists.

1.3 Higher-order assessment rules

Panchanathan and Boyd (2003) also showed that Sugden's (1986) *standing* rule can be an evolutionarily stable strategy (ESS) in this model, as had previously been suggested by Leimar and Hammerstein (2001). But standing, unlike image scoring, is a second-order assessment rule, since in updating an individual's reputation after observing its action it takes into account the reputation of the individual's opponent. This allows it to distinguish "justified" and "unjustified" defections.[3] Later literature has almost exclusively focused on higher-order assessment rules, see the survey of Sigmund (2012). In the last decade the overall picture has emerged that higher-order assessment rules can stabilize reputation-based cooperation, while first-order assessment rules cannot.[4] But higher-order assessment rules are cognitively highly demanding. Moreover, they heavily rely on the reputations of individuals being built and truthfully spread by word-of-mouth by observers. Their superiority therefore received a blow when Suzuki and Kimura (2013) showed that introducing arbitrarily small costs for building or spreading a reputation results in a second-order social dilemma and renders cooperation impossible.

All in all, the situation seems puzzling: Higher-order assessment rules are cognitively highly demanding and work only under the unrealistic assumption of costless reputation transmission. The first-order assessment rule of image scoring, on the other hand, renders reputation-based cooperation invadable by unconditional cooperators[5] in the absence of noise and unstable in the presence of noise in the binary case. But indirectly reciprocal behavior in humans, and especially image scoring is strongly supported by experimental research (Wedekind and Milinski, 2000, Milinski et al, 2001, Bolton et al, 2005, Seinen and Schram, 2006, Engelmann and Fischbacher, 2009). How can the prevalence of indirect reciprocity be reconciled with the fragility of its theoretical foundations?

1.4 Beyond binary scores

Our answer in this paper is that within the space of threshold strategies, instability of cooperation under image scoring in the presence of noise obtains *only* for the binary scoring case and not for the general case of image scoring. We present a simplified version of the original "full score" model of image scoring (Nowak and Sigmund, 1998a),

[3] Standing and a range of other sophisticated higher-order assessment rules can successfully stabilize cooperation based on indirect reciprocity, as has later been shown by Ohtsuki (2004), Ohtsuki and Iwasa (2004, 2006), and Brandt and Sigmund (2004). This literature is reviewed in Nowak and Sigmund (2005).

[4] There are a few exceptions, but these are based on rather special assumptions like a fixed or Poisson-distributed number of perfectly synchronized rounds of interaction (Fishman, 2003, Brandt and Sigmund, 2004), monotonically growing social networks (Brandt and Sigmund, 2005), or interactions in larger groups (Suzuki and Akiyama, 2007, 2008).

[5] Note that this is disastrous for cooperation, since unconditional cooperators are easy prey for defectors. This means that not only is reputation-based cooperation not evolutionarily stable, but it is not even robust against indirect invasions (van Veelen, 2012).

where an individual observes several past actions of its partner. Following the original model, we also assume that individuals condition their choice of action on the partner's image score using a threshold strategy, meaning that they cooperate if and only if the partner's perceived score exceeds some threshold. We then introduce noise in the form of implementation errors and show that in this model, reputation-based cooperation is indeed evolutionarily stable under a wide range of parameter values.

Here, a note on the set of feasible strategies we consider is in order. Heller and Mohlin (2015) demonstrated that if individuals observe only partners' past actions and the set of feasible strategies is completely unrestricted, then defection is the only stable outcome. Intuitively, under our assumptions an individual's payoff is only determined by its past cooperation rate. Thus, in any population there exists an optimal (i.e. payoff-maximizing) individual cooperation rate, and if this rate is strictly positive, non-discriminatory mutants randomly cooperating with the optimal probability can invade any such population. When we talk of evolutionary stability, we therefore follow the majority of the literature by referring to stability within a restricted strategy space, given by the set of feasible strategies allowed by the model. In general, limitations of feasible strategies are due to biological constraints on cognitive capacity, memory length, or physiology (Broom and Rychtar, 2013).

1.5 Threshold strategies

The image scoring literature has typically studied rather small strategy spaces. For example, in the analytical part (on binary scoring) of their work, Nowak and Sigmund (1998a, 1998b) give conditions for which discriminators are evolutionarily stable, but there they consider only two strategies, viz. defectors and discriminators. Similarly, Brandt and Sigmund (2004, 2005) consider asymptotic stability under the replicator dynamics (which is implied by evolutionary stability) in a three-strategies model of defectors, cooperators, and discriminators. Berger (2011) studies stability in a ternary scoring model comprising defectors, cooperators, and "tolerant" discriminators. In contrast, here we study a multiple scoring model based on observations of n past actions of one's current partner, allowing for *all* $n + 2$ threshold strategies, including the unconditional extremes of always defecting and always cooperating. We thus consider a setting which closely resembles the original "full score" model of Nowak and Sigmund (1998a).

Threshold strategies have been introduced by Taylor (1976) in the context of an iterated N-person Prisoner's Dilemma game as a natural generalization of the well-known Tit for Tat strategy for the two-person case. These strategies can be coded by a single number, and are therefore the simplest nonconstant strategies in this context. The restriction to threshold strategies has also been used for the multiplayer Prisoner's Dilemma and the binary public goods game by Boyd and Richerson (1988), Kurokawa and Ihara (2009), and van Segbroeck et al. (2012), amongst others, and, as mentioned, by Nowak and Sigmund (1998a) for their study of indirect reciprocity.

We show that introducing noise in the form of unintended defections may stabilize cooperation provided $n \geq 2$. This excludes precisely—and only—the binary scoring case $n = 1$ which was the basis of almost all previous analytical studies of image scoring. Our main result on the existence of cooperative ESS depends on the assumption that individuals play deterministic strategies, but not on the assumption that only threshold strategies are feasible. Generically, our discriminator ESS are also immune against invasion of mutants playing non-threshold strategies.

Our stability results are actually slightly stronger than evolutionary stability. We show that the ESS we find are strict or quasi-strict Nash equilibria. By continuity this implies that they remain stable even if we introduce small costs of observation. Thus, our results are immune against the critique of Suzuki and Kimura (2013) discussed above.

In our model, members of a stable discriminator population cooperate with each other with a high rate, which is, however, smaller than the maximal possible rate given by the complementary of the error rate. Mutants with a lower tolerance level (defection threshold) than the incumbent cooperate less often and therefore receive considerably less help by incumbents, which pushes the mutant's payoff below the incumbent's. On the other hand, mutants with a higher tolerance level cooperate more frequently, paying higher costs but also being helped more often by incumbents. However, the increase in benefits received from incumbents is rather small and cannot offset the increased costs. Thus, also the more tolerant mutants cannot invade. This kind of deterrence of mutants with lower as well as higher threshold levels can only arise since the functions mapping mutants' past cooperation rates to incumbents' cooperation probability against mutants are nonlinear.

2 Model

2.1 The donation game, errors, and threshold strategies

Consider an infinite population of individuals. Time t is continuous and pairs of individuals are repeatedly and randomly drawn to interact in a donation game. During each interaction, one individual is randomly chosen to be the donor and the other to be the receiver. Donors can either give help (cooperate, C) or not (defect, D) to the receiver. Helping decreases the donor's payoff by an amount c and increases the receiver's payoff by b, where $b > c > 0$. For convenience we will make the usual assumption that actually each individual plays in both roles at the same time during an interaction.[6] With a small probability $\alpha > 0$ a donor who intends to cooperate is not able to do so (e.g. due to lack of resources) and instead defects. No implementation errors are assumed if a

[6]This means that on each interaction, individuals play a Prisoner's Dilemma with *equal gains from switching*, also called a *linear* Prisonser's Dilemma.

donor intends to defect.[7]

Before a donor implements his action, he is informed of his partner's choices in a random sample of $n \geq 1$ past interactions of this partner.[8] The donor's action then depends on the donor's strategy and on the number of defections (D's) in the drawn sample. A donor with a threshold-i strategy intends to cooperate if and only if his partner defected at most i times in the sample. The threshold level i is thus a measure of the donor's tolerance against defections. An individual playing this strategy is called an *i-discriminator*. We let $-1 \leq i \leq n$ to include the unconditional strategies ALLC ($i = n$) and ALLD ($i = -1$). We employ the polymorphic interpretation of mixed strategies, i.e. we assume that individuals play pure strategies, while mixed population states represent polymorphic states.

Our stability analysis relies on the static ESS concept. However, payoffs of different discriminator strategies depend on their cooperation rates. While cooperation rates change dynamically, we assume that interactions are much more frequent than reproduction. Technically we assume that cooperation rates have already converged to some fixed point before reproduction occurs, allowing us to calculate payoffs in the stationary states of cooperation rates. Convergence from arbitrary initial values is shown in section 2.3 below.

2.2 Cooperation functions

Assume that an i-discriminator meets an individual with a past frequency of cooperation given by p. Then the probability that the i-discriminator helps this individual is a function of p only. We call this the *cooperation function* of the i-discriminator and denote it by $f_i(p)$. From our assumptions it follows that

$$\begin{aligned} f_{-1}(p) &\equiv 0, \\ f_i(p) &= (1-\alpha)F(i; n, 1-p) \text{ for } i \in \{0, \ldots, n\} \end{aligned} \tag{1}$$

where $F(i; n, 1-p)$ denotes the cumulative distribution function of the binomial distribution, i.e., the probability that in an n-times repeated Bernoulli experiment with probability $1-p$ of outcome D in one experiment, the D appears at most i times. The case $n = 5$ and $\alpha = 0.1$ is displayed in Figure 1.

[7] This assumption is standard in the literature. First, since in the donation game defecting basically just means not to act, "unintended cooperation" would be an implausible assumption. Second, this asymmetry in the noise makes the case against cooperation stronger and therefore strengthens the model's results.

[8] Note that we do not assume that these are the partner's last n actions. So this assumption cannot be formulated in terms of "memory length".

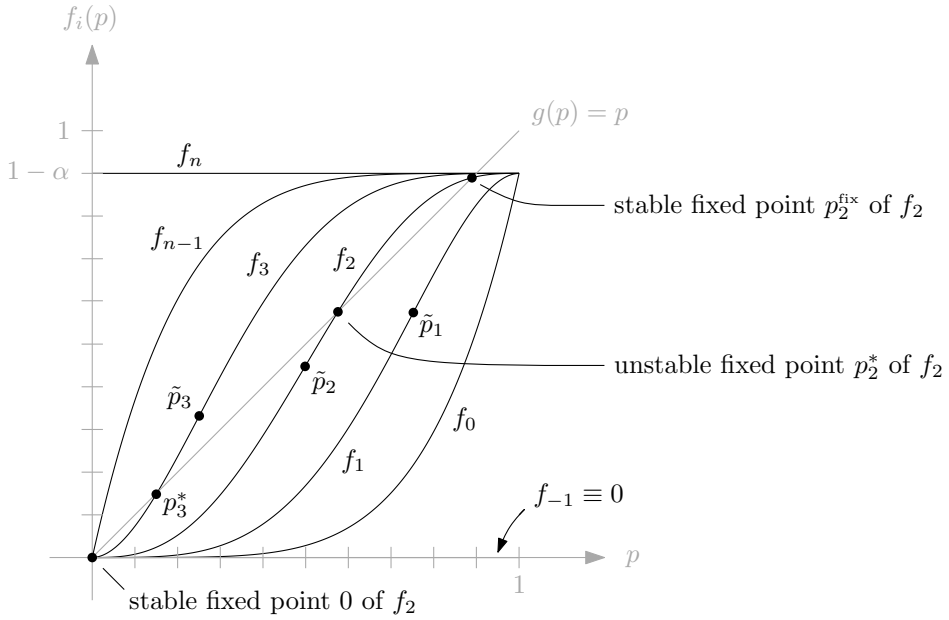

Figure 1: Cooperation functions $f_i(p)$ for $n = 5$ and $\alpha = 0.1$.

For the special cases $i = -1$ and $i = n$ we have the constant cooperation functions

$$f_{-1}(p) \equiv 0 \quad \text{(ALLD)} \qquad \text{and} \qquad f_n(p) \equiv 1 - \alpha \quad \text{(ALLC)}. \tag{2}$$

From now on, in this subsection, we restrict our attention to the cooperation functions of proper discriminators, i.e. $0 \le i \le n-1$. Writing the binomial distribution function as a regularized beta function we obtain

$$\begin{aligned} f_i(p) &= (1-\alpha)\sum_{k=0}^{i}\binom{n}{k}p^{n-k}(1-p)^k \\ &= (1-\alpha)(n-i)\binom{n}{i}\int_0^p s^{n-i-1}(1-s)^i\,\mathrm{d}s \end{aligned} \tag{3}$$

The cooperation functions of proper discriminators are strictly increasing from $f_i(0) = 0$ to $f_i(1) = 1 - \alpha$.

Two important special cases are

$$f_0(p) = (1-\alpha)p^n \qquad \text{and} \qquad f_{n-1}(p) = (1-\alpha)(1-(1-p)^n). \tag{4}$$

Using the identity provided by the beta function we can calculate the derivatives

$$f_i'(p) = (1-\alpha)(n-i)\binom{n}{i}p^{n-i-1}(1-p)^i \tag{5}$$

These are non-negative and vanish at $p = 0$ (except for $i = n - 1$) and at $p = 1$ (except for $i = 0$).

The second derivatives for $i \in \{0, \ldots, n-1\}$ are given by

$$f_i''(p) = (1-\alpha)(n-i)\binom{n}{i}p^{n-i-2}(1-p)^{i-1}(n-i-1-(n-1)p). \tag{6}$$

In particular,

$$\begin{aligned} f_0''(p) &= (1-\alpha)n(n-1)p^{n-2}, \\ f_{n-1}''(p) &= -(1-\alpha)n(n-1)(1-p)^{n-2}. \end{aligned}$$

For $n = 1$ we have $f_0(p) = (1-\alpha)p$, $f_0'(p) \equiv 1-\alpha$, and $f_0''(p) \equiv 0$. For $n \geq 2$ we can see that for every $i \in \{0, \ldots, n-1\}$, $f_i'(.)$ strictly increases from $p = 0$ up to the inflection point

$$\tilde{p}_i = \frac{n-i-1}{n-1} \tag{7}$$

and then strictly decreases until $p = 1$. In other words, $f_i(.)$ is strictly convex on $[0, \tilde{p}_i]$ and strictly concave on $[\tilde{p}_i, 1]$. Note that for the special case $i = 0$, we have $\tilde{p}_0 = 1$, so $f_0(.)$ is strictly convex on $[0, 1]$. Analogously, $\tilde{p}_{n-1} = 0$ and $f_{n-1}(.)$ is strictly concave on $[0, 1]$.

2.3 Cooperation rate dynamics and fixed points

Consider now, for $i \in \{0, \ldots, n-1\}$, a homogeneous population of i-discriminators. Assume that at time t the past cooperation rate in the population is $p(t)$. If in a small time interval Δ a proportion Δ of the population is chosen to play, then a fraction $f_i(p(t))$ of those individuals will cooperate, so $p(t+\Delta) = [tp(t) + \Delta f_i(p(t))](t+\Delta)^{-1}$. Letting $\Delta \to 0$ we arrive at the cooperation rate dynamics $\dot{p} = [f_i(p) - p]t^{-1}$. Hence, as long as $f_i(p(t)) < p(t)$, the cooperation rate will decrease, and as long as $f_i(p(t)) > p(t)$, the cooperation rate will increase. The cooperation rate will thus converge to a fixed point of the cooperation function f_i.[9]

The special case $n = 1$, where only a single past action is observed, leads to the binary scoring model. Note that since $f_0(p) = (1-\alpha)p < p$, convergence of cooperation rates to 0 is inevitable. Discrimination based on single observations does not work. For $n = 1$, a homogeneous population of discriminators always ends up with pure defection in the

[9] Since the population is large, partners are matched randomly and partners' past actions are drawn randomly, each individual samples independently from the same distribution of Cs and Ds with mean $p(t)$. From the martingale central limit theorem it follows that individuals' cooperation rates converge pairwise in probability. Therefore the overall cooperation rate converges as well.

long run. However, as we demonstrate below, this result does not extend to the general case of $n \geq 1$.

In Figure 1 the fixed points of the cooperation function f_i are the intersections of f_i with the diagonal. For small i, $p = 0$ is the unique fixed point and the population ends up with all-out defection. This is always the case for $i = -1$ and $i = 0$, but it might also hold for larger values of i, if the error rate α is large. But if α is small enough, then for some minimal i-value another stable fixed point $\tilde{p}_i < p_i^{\text{fix}} \leq 1 - \alpha$ appears on the concave part of the graph of the i-discriminator's cooperation function, accompanied by an unstable fixed point $0 \leq p_i^* < p_i^{\text{fix}}$. This is the case whenever the cooperation function f_i crosses the diagonal from above.

So, generically, for given α, $n \geq 2$, and $-1 \leq i \leq n$, we either have a unique and globally attracting fixed point at 0 (all-out defection), a bistable situation with either all-out defection or a high cooperation rate in the long run, or—for $i = n - 1$, where 0 is an unstable fixed point, and for the unconditional cooperators $i = n$—a highly cooperative population in the long run. The latter two cases are those where a homogeneous population of i-discriminators is able to maintain a high rate of cooperation.[10] We then say that the i-discriminators are *self-cooperative*. Technically, i-discriminators are self-cooperative if and only if their cooperation function f_i crosses the diagonal from above.

Self-cooperation is always obtained for the case $i = n$ (an ALLC-population), and from (5) also for $i = n - 1$ provided $\alpha < \frac{n-1}{n}$. However, for $i \leq n - 2$ self-cooperation is only possible in the bistable case. The dynamics of cooperation rates then allow for a cooperative as well as for a defective regime, depending on initial conditions. Hence, to uniquely determine the final cooperation rate of the population we have to make an assumption on those initial conditions. We assume here that newborn individuals, who lack a record of past play, are given the benefit of doubt, i.e. they are treated by discriminators as if they had a clean record of all-out cooperation. It then follows that self-cooperative discriminators always end up with a cooperation rate at the high-cooperation fixed point p_i^{fix}.[11]

2.4 Payoffs

Let us now investigate whether a small fraction of mutant m-discriminators can survive or even spread in an otherwise homogeneous incumbent population of self-cooperative i-discriminators. In any such investigation we will assume that prior to the mutant's

[10] Note that "high" is to be understood here as relative to the maximum possible cooperation rate of $1 - \alpha$.

[11] This assumption is not crucial, however. Without it, our results would continue to hold provided the population happens to start with a cooperation rate within the basin of attraction of the high-cooperation fixed point.

entry the incumbent's cooperation rate has already stabilized at p_i^{fix}. We denote the expected payoff per interaction of a single j-discriminator in an otherwise homogeneous population of i-discriminators by $\hat{\pi}(j|i)$. It is useful to work with normalized payoffs, measuring original payoffs in multiples of the benefit b, so let $\pi(j|i) := b^{-1}\hat{\pi}(j|i)$.[12]

When a mutant m-discriminator enters an incumbent population of i-discriminators, the incumbents' overall cooperation rate remains at p_i^{fix}, which implies a mutant's cooperation rate of $f_m(p_i^{\text{fix}})$. Hence, upon meeting the mutant, an incumbent will cooperate with probability $f_i(f_m(p_i^{\text{fix}}))$.

For the mutant's payoff we thus get $\hat{\pi}(m|i) = bf_i(f_m(p_i^{\text{fix}})) - cf_m(p_i^{\text{fix}})$, or

$$\pi(m|i) = f_i(f_m(p_i^{\text{fix}})) - rf_m(p_i^{\text{fix}}), \tag{8}$$

where $r := c/b$ denotes the cost-benefit ratio of the donation.

For an incumbent, the probability of meeting the mutant is negligible, so the incumbents' average payoff will be $\pi(i|i) = (1-r)p_i^{\text{fix}}$.

2.5 Evolutionary stability of discrimination

A sufficient condition for the incumbent population to be evolutionarily stable in the sense of Maynard Smith and Price (1973) is that the incumbent's payoff is strictly larger than any mutant's payoff, i.e. that $\pi(i|i) > \pi(m|i)$, or

$$(1-r)p_i^{\text{fix}} > f_i(f_m(p_i^{\text{fix}})) - rf_m(p_i^{\text{fix}}) \tag{9}$$

for all $m \neq i$.

It is easy to see that for any n, unconditional cooperators, i.e. n-discriminators, can always be invaded by unconditional defectors. Strictly speaking, defectors themselves are not evolutionarily stable, because mutant discriminators do not cooperate with them, earn 0 payoff as well and can grow by neutral drift. However, these mutants never manage to cooperate with each other, since their cooperation rate, having started at $p_i = 0$, never reaches the basin of attraction $p_i > p_i^*$ of the cooperative regime. So even if ALLD is not evolutionarily stable, defection cannot be overcome.[13] Essentially the same is true for the 0-discriminator, which is never self-cooperative. Hence ESS candidates exist only for $n \geq 2$ and $1 \leq i \leq n-1$.

So let us assume that $n \geq 2$ and α and $1 \leq i \leq n-1$ are such that the i-discriminator is self-cooperative with cooperation rate p_i^{fix}. Self-cooperativeness implies that at p_i^{fix} the

[12] The ESS concept is immune to rescaling of payoffs, so normalizing payoffs is without loss of generality here (Berger, 2009).

[13] Taking into account our assumption that proper discriminators cooperate with newborn defectors, defectors even have a slight advantage, making ALLD evolutionarily stable. However, this assumption is extraneous to the model.

cooperation function f_i crosses the diagonal from above, i.e. $f_i'(p_i^{\text{fix}}) < 1$. Moreover, the graph of f_i is below the diagonal between 0 and p_i^*, has slope greater than 1 between p_i^* and $\tilde{p}_i$, and is strictly concave between $\tilde{p}_i$ and 1. This implies that the graph of f_i is completely below the tangent to f_i at p_i^{fix}. Applying this at the point $p = f_m(p_i^{\text{fix}})$ we get the inequality $f_i(f_m(p_i^{\text{fix}})) < p_i^{\text{fix}} - f_i'(p_i^{\text{fix}})[p_i^{\text{fix}} - f_m(p_i^{\text{fix}})]$. Assume now that the cost-benefit ratio r happens to be exactly equal to $r = f_i'(p_i^{\text{fix}})$, then the inequality can be written as $f_i(f_m(p_i^{\text{fix}})) < (1-r)p_i^{\text{fix}} + r f_m(p_i^{\text{fix}})$. Comparing this to inequality (9) shows that this means $\pi(i|i) > \pi(m|i)$, implying evolutionary stability of the incumbent i-discriminator. By continuity of both sides of the inequality in r, evolutionary stability continues to hold for nearby cost-benefit ratios. This proves:

Theorem 1 (Existence of ESS-discriminators)**.** *Fix $n \geq 2$ and $0 < \alpha < \frac{n-1}{n}$. Choose $1 \leq i \leq n-1$ such that the i-discriminator is self-cooperative. Then there exists an open interval of cost-benefit ratios r such that the i-discriminator is evolutionarily stable.*

Intuitively, in a population of self-cooperative i-discriminators a mutant strategy's payoff depends only on its cooperation rate and hence there will be some optimal cooperation rate. But the peculiar shape of the incumbent's cooperation function implies that by varying r between 0 and 1 we can always find a cost-benefit ratio such that this optimal cooperation rate equals the incumbent's own cooperation rate p_i^{fix}. Any possible mutant will then achieve a strictly lower payoff against the incumbent than the incumbent itself. By continuity of payoffs in the cost-benefit ratio, this continues to hold for nearby such ratios.

If the i-discriminator is evolutionarily stable, a homogeneous population of i-discriminators cooperates at a high rate and resists invasion attempts of all mutant m-discriminators, including ALLC and ALLD. If the error rate α is small enough, all i-discriminators with $1 \leq i \leq n-1$ are self-cooperative and hence each i-discriminator is an ESS for some open set of cost-benefit ratios. The only case where no such ESS exists is the binary image scoring case, i.e. $n = 1$, where the only proper discriminator, $i = 0$, is not self-cooperative for any $\alpha > 0$.

Note that in the argument leading to Theorem 1 we did not make use of the assumption that a mutant employs a threshold strategy. Indeed, the result continues to hold whenever for the chosen values of α and n there is no mutant strategy m such that $f_m(p_i^{\text{fix}}) = p_i^{\text{fix}}$, i.e. whenever no mutant's cooperation function intersects the incumbent's cooperation function exactly on the diagonal. However, since there are only finitely many possible mutants and all cooperation functions are scaled by the factor $1-\alpha$ this happens only in non-generic cases. Theorem 1 can therefore be extended in the following way.

Proposition 1. *Fix $n \geq 2$ and let the strategy space include all 2^{n+1} strategies corresponding to mappings from $\{0, 1, 2, \ldots, n\}$ to $\{C,D\}$. Let $0 < \alpha < \frac{n-1}{n}$ and choose $1 \leq i \leq n-1$ such that the i-discriminator is self-cooperative. Then, generically, there*

exists an open interval of cost-benefit ratios r such that the i-discriminator is evolutionarily stable.

3 ESS Regions

3.1 Overview

The exact shape of the ESS-regions R_i in the interior of the α-r-square where an i-discriminator is an ESS, can be determined numerically from inequality (9). It turns out that for small α the open intervals of r-values guaranteeing the ESS-property for the i-discriminators can be extremely small. However, a sizable fraction of the α-r-square consists of parameter combinations where some discriminator is an ESS. For $n = 5$ these ESS regions are depicted in Figure 2 as the "leaves" originating from $(0, 0)$.

Note that relatively large values of $\alpha < 1$ cannot readily be interpreted as probabilities of implementation errors of intended donations. Rather, high values of α indicate that individuals intending to help often simply lack the resources to do so.[14] For very large values of α, viz. $\alpha \geq \frac{n-1}{n}$, not even the most tolerant discriminator $i = n - 1$ is self-cooperative, and cooperation is doomed to fail. However, as can be seen from region R_4 in Figure 2, for medium to high cost-benefit ratios the most tolerant discriminator remains an ESS even for α-values arbitrarily close to the maximum of 0.8 for $n = 5$ (this is proved rigorously below). Clearly, however, the cooperation rate in this "cooperative" regime is actually rather low, being bounded from above by the corresponding $(1 - \alpha)$-values close to 0.2.

3.2 Properties of ESS regions

The numerical calculations behind Figure 2 suggest that for every i-discriminator with $1 \leq i \leq n - 1$ there exists a certain $\alpha_i^{\max}$ such that the discriminator can be an ESS for all $0 < \alpha < \alpha_i^{\max}$ but is never an ESS for $\alpha > \alpha_i^{\max}$. This is indeed the case. The proof of the existence of ESS discriminators above is valid as long as the discriminator in question is self-cooperative. For $1 \leq i \leq n-1$ this is the case whenever α is small enough. Increasing α scales down the cooperation functions in Figure 1 until the unstable fixed point p_i^* and the stable fixed point p_i^{fix} coincide and the diagonal is tangential to the cooperation function at this value. The value of α where this happens is $\alpha_i^{\max}$. From Figure 1 it is also immediate that $\alpha_i^{\max}$ is increasing in i.

A special case is $\alpha_{n-1}^{\max}$, where self-cooperativeness of the most tolerant proper discriminator breaks down. Since f_{n-1} is strictly concave, a stable fixed point $p_{n-1}^{\text{fix}} > 0$ exists

[14] This could suggest an interpretation of such a high-α population as a poor society.

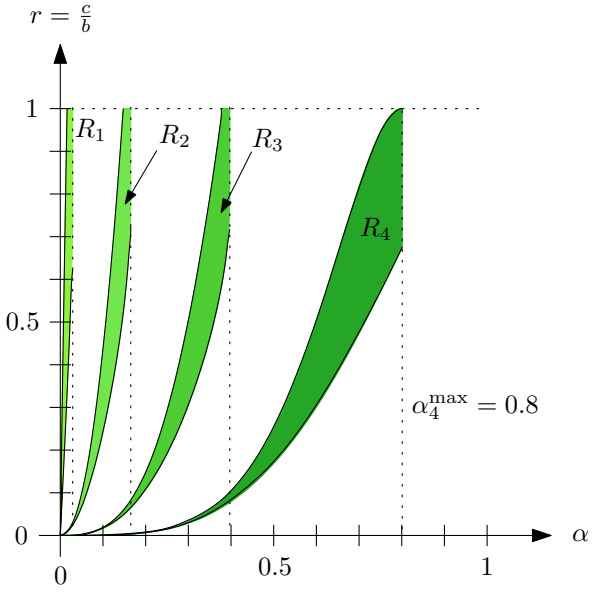

Figure 2: ESS regions $R_1, \ldots, R_4$ for $n = 5$.

if and only if $f'_{n-1}(0) > 1$. From equation (5) we have $f'_{n-1}(p) = (1-\alpha)n(1-p)^{n-1}$, hence $f'_{n-1}(0) = (1-\alpha)n$, implying $\alpha_{n-1}^{\max} = 1 - \frac{1}{n}$.

We have shown that in the case of self-cooperation, i.e. for $\alpha < \alpha_i^{\max}$, there exists an open interval of cost-benefit ratios r such that the i-discriminator is evolutionarily stable. By construction, this interval contains the ratio $r = f'_i(p_i^{\text{fix}})$. Maximally extending the boundaries of the interval leads to the largest such interval $r_i^{\min} < r < r_i^{\max}$. Note that the boundary values depend on α. Given any cooperation function f, let us now denote the slope of the line between the two points $(p_1, f(p_1))$ and $(p_2, f(p_2))$ on the graph of f by $\text{sl}_f(p_1, p_2) := \frac{f(p_2)-f(p_1)}{p_2-p_1}$. We can then show that $r_i^{\min} = \text{sl}_{f_i}(p_i^{\text{fix}}, f_{i+1}(p_i^{\text{fix}}))$ and $r_i^{\max} = \min\left(1, \text{sl}_{f_i}(f_{i-1}(p_i^{\text{fix}}), p_i^{\text{fix}})\right)$. Moreover, if the cost-benefit ratio is outside the closure of this interval, the i-discriminator can be invaded by a mutant strategy and is never an ESS. The proof of this is relegated to the Appendix.

Figure 2 also strongly suggests that the ESS regions of different i-discriminators do not overlap. The ESS regions of more tolerant discriminators seem to lie to the right and below the ones of stricter discriminators. Indeed, this is the case. Again the proof can be found in the Appendix.

3.3 Evolutionarily stable mixtures of neighboring discriminators

Figure 2 raises one more question. What happens in the regions where no i-discriminator is an ESS? We try to answer this question in this section.

First we focus on points in parameter space which lie vertically between the ESS regions of an i-discriminator and the $(i+1)$-discriminator. These are points (α, r) such that there exists an $i \in \{1, \ldots, n-1\}$ with $\alpha < \alpha_i^{\max}$ and $r_{i+1}^{\max} < r < r_i^{\min}$. We show that in any such case there exists a mixture of i- and $(i+1)$-discriminators which cannot be invaded by any mutant strategy and thus is an evolutionarily stable state.

If i- and $(i+1)$-discriminators are present in fixed proportions in a well-mixed population, the dynamics of their respective cooperation rates $(p_i, p_{i+1})(t)$ are described by a smooth two-dimensional dynamical system. It is easy to see that if the initial cooperation rates of both groups are close to zero, they both vanish in the limit. However, we show now that there is always a second asymptotically stable fixed point with high cooperation rates. As in the case of a single type of discriminators, we will assume that initial cooperation rates are high, which allows us to treat them as fixed at their respective equilibrium values with high cooperation when looking for evolutionarily stable mixtures of i- and $(i+1)$-discriminators.

3.3.1 Limit cooperation rates in mixtures of two discriminators

Assume the population is composed of a fraction q of i-discriminators and a fraction $1-q$ of $(i+1)$-discriminators. Let the initial past cooperation rates in the two groups be p_i and p_{i+1}. On his next interaction, an i-discriminator meets another i-discriminator with probability q and an $(i+1)$-discriminator with probability $1-q$, in the first case cooperating with probability $f_i(p_i)$ and in the second case cooperating with probability $f_i(p_{i+1})$, so in his next interaction his cooperation probability will be $qf_i(p_i) + (1-q)f_i(p_{i+1})$. The cooperation rate of the i-discriminators will thus be moved into this direction. The analogous applies to the $(i+1)$-discriminator. Dropping the common factor t^{-1}, the dynamics of cooperation rates can hance be described by

$$\begin{aligned} \dot{p}_i &= qf_i(p_i) + (1-q)f_i(p_{i+1}) - p_i \\ \dot{p}_{i+1} &= qf_{i+1}(p_i) + (1-q)f_{i+1}(p_{i+1}) - p_{i+1} \end{aligned} \tag{10}$$

Consider now the case where $p_i = p_i^{\text{fix}}$ and $p_{i+1} > p_i$. Since f_i is strictly increasing, $f_i(p_{i+1}) > f_i(p_i) = p_i$ and the first equation in (10) implies that $\dot{p}_i > 0$. Analogously, $p_{i+1} = p_{i+1}^{\text{fix}}$ and $p_{i+1} > p_i$ imply $\dot{p}_{i+1} < 0$. On the other hand, since f_{i+1} is strictly greater than f_i on the interior of the unit interval, subtracting the first from the second equation of (10) implies $\frac{\mathrm{d}}{\mathrm{d}t}(p_{i+1} - p_i) > -(p_{i+1} - p_i)$, so $p_{i+1} - p_i$ strictly increases at interior points of the diagonal $\{p_i = p_{i+1}\}$. Hence, the triangle $\{p_{i+1}^{\text{fix}} \geq p_{i+1} \geq p_i \geq p_i^{\text{fix}}\}$ in phase space is forward invariant.

Consider next the isocline $\dot{p}_i = 0$ in this triangle. We have $\dot{p}_i = 0$ at the lower left corner $(p_i^{\text{fix}}, p_i^{\text{fix}})$, $\dot{p}_i > 0$ at the upper left corner $(p_i^{\text{fix}}, p_{i+1}^{\text{fix}})$, and $\dot{p}_i < 0$ at the upper right

corner $(p_{i+1}^{\text{fix}}, p_{i+1}^{\text{fix}})$, so the isocline runs from the lower left corner to the upper edge of the triangle. Since $\frac{\partial}{\partial p_{i+1}}[qf_i(p_i) + (1-q)f_i(p_{i+1}) - p_i] = (1-q)f_i'(p_{i+1}) > 0$, the implicit function theorem tells us that the isocline $\dot{p}_i = 0$ can be written as a function $p_{i+1}(p_i)$ with $p_{i+1}'(p_i) = -\frac{qf_i'(p_i)-1}{(1-q)f_i'(p_{i+1})}$. Since $f_i'(p_i) < 1$, we have $p_{i+1}'(p_i) > 0$. From this we get $p_{i+1}''(p_i) = -\frac{qf_i''(p_i)(1-q)f_i'(p_{i+1}) - (qf_i'(p_i)-1)(1-q)f_i''(p_{i+1})p_{i+1}'(p_i)}{[(1-q)f_i'(p_{i+1})]^2} > 0$. So the isocline $\dot{p}_i = 0$ can be written as an increasing and convex function running from the lower left corner to the upper edge of the triangle. The analogous arguments for p_{i+1} show that the isocline $\dot{p}_{i+1} = 0$ can be written as an increasing and convex function running from the left edge to the upper right corner of the triangle. By continuity of these functions they intersect in a unique fixed point in the interior of the triangle, which we denote by $(p_{i,i+1}^{\text{fix}}, p_{i+1,i}^{\text{fix}})$. This fixed point has $p_i^{\text{fix}} < p_{i,i+1}^{\text{fix}} < p_{i+1,i}^{\text{fix}} < p_{i+1}^{\text{fix}}$ and is asymptotically stable. [15]

3.3.2 Evolutionary stability of mixtures of two discriminators

Let $q \in [0,1]$ be fixed. Consider again the population mixture of a fraction q of i-discriminators and a fraction $1-q$ of $(i+1)$-discriminators. As shown above, the cooperation rates of the two groups will then equilibrate at $p_{i,i+1}^{\text{fix}}$ and $p_{i+1,i}^{\text{fix}}$, respectively. Therefore, the payoffs of an i-discriminator and an $(i+1)$-discriminator are given by

$$\begin{aligned} \pi_i &= qf_i(p_{i,i+1}^{\text{fix}}) + (1-q)f_{i+1}(p_{i,i+1}^{\text{fix}}) - rp_{i,i+1}^{\text{fix}}, \\ \pi_{i+1} &= qf_i(p_{i+1,i}^{\text{fix}}) + (1-q)f_{i+1}(p_{i+1,i}^{\text{fix}}) - rp_{i+1,i}^{\text{fix}}, \end{aligned} \tag{11}$$

respectively. We now define a new function, which is just a weighted average of f_i and f_{i+1}, viz.

$$f_{i,i+1}^q(p) := qf_i(p) + (1-q)f_{i+1}(p)$$

The two payoffs can then be written as $\pi_i = f_{i,i+1}^q(p_{i,i+1}^{\text{fix}}) - rp_{i,i+1}^{\text{fix}}$ and $\pi_{i+1} = f_{i,i+1}^q(p_{i+1,i}^{\text{fix}}) - rp_{i+1,i}^{\text{fix}}$. Division by $p_{i+1,i}^{\text{fix}} - p_{i,i+1}^{\text{fix}}$ shows that the payoff difference $\pi_i - \pi_{i+1}$ has the same sign as the difference between the cost-benefit ratio r and the slope of the line connecting the two points $(p_{i+1,i}^{\text{fix}}, f_{i,i+1}^q(p_{i+1,i}^{\text{fix}}))$ and $(p_{i,i+1}^{\text{fix}}, f_{i,i+1}^q(p_{i,i+1}^{\text{fix}}))$, i.e. the difference $r - \text{sl}_{f_{i,i+1}^q}(p_{i,i+1}^{\text{fix}}, p_{i+1,i}^{\text{fix}})$. In particular, equality of payoffs implies $\text{sl}_{f_{i,i+1}^q}(p_{i,i+1}^{\text{fix}}, p_{i+1,i}^{\text{fix}}) = r$.

Note that $q = 0$ is just the situation of a homogeneous population of $(i+1)$-discriminators, and $\text{sl}_{f_{i,i+1}^q}(p_{i,i+1}^{\text{fix}}, p_{i+1,i}^{\text{fix}}) = \text{sl}_{f_{i+1}}(f_i(p_{i+1}^{\text{fix}}), p_{i+1}^{\text{fix}}) = r_{i+1}^{\max}$. By our assumption of $r_{i+1}^{\max} < r < r_i^{\min}$ we have $\pi_i - \pi_{i+1} > 0$, so this population can be invaded by i-discriminators.

[15] Indeed it can be shown that this fixed point attracts all solutions of (10) with initial cooperation rates exceeding p_i^*.

Vice versa, for $q = 1$ we get a homogeneous population of i-discriminators, which can be invaded by $(i+1)$-discriminators. By continuity of the payoffs in q, there must exist a $0 < q < 1$ such that in the resulting mixture of i- and $(i+1)$-discriminators, both groups have equal payoffs. This mixture can neither be invaded by i- nor by $(i+1)$-discriminators. It remains to be shown that also no other mutant m-discriminator can invade.

Theorem 2 (Stable mix of two discriminators)**.** *Let $i \in \{1, \ldots, n-1\}$, $0 < \alpha < \alpha_i^{\max}$, and $r_{i+1}^{\max} < r < r_i^{\min}$. Then there exists a unique mixture of i- and $(i+1)$-discriminators which is an ESS.*

The proof of Theorem 2 can be found in the Appendix. Figure 3 shows the ESS-regions of mixtures of neighboring discriminators added to the ESS-regions of single discriminators in the α-r-square.

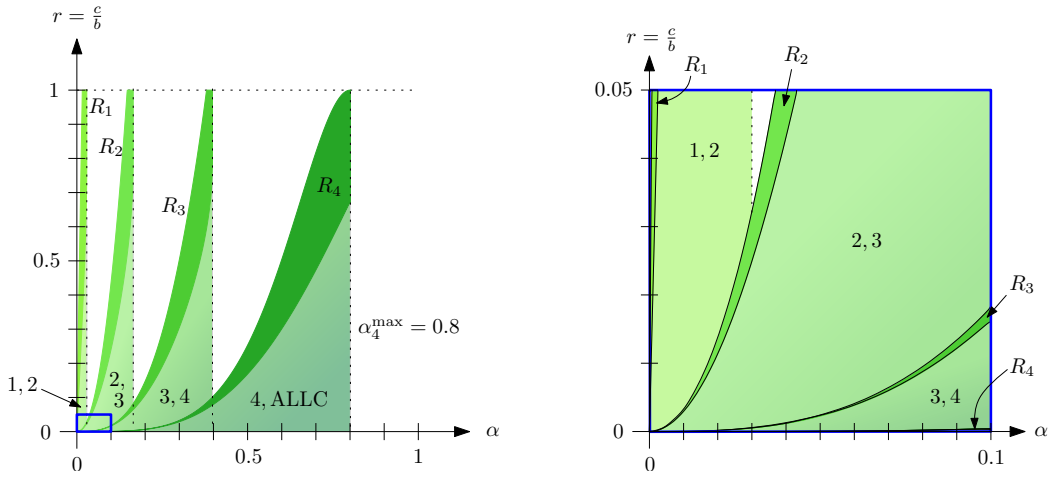

Figure 3: ESS regions of single discriminating strategies and mixtures of two discriminating strategies for $n = 5$. On the right side a zoom of the area close to the origin.

3.4 The chances for evolutionary stability

Given randomly selected values for α and r in the unit interval, what is the probability that a high rate of cooperation can be achieved in an ESS? We can't answer this question exactly, since we have only proved existence of two special types of ESS here, single discriminator ESS and ESS of mixtures of two neighboring discriminators. It is in principle possible that some fraction of the white region in Figure 3 could admit similar or other types of ESS. However, by measuring the coloured area in this Figure we can at least calculate a lower bound for the chances that a cooperative ESS exists. Table 1 provides these values for realistically low as well as for intermediate and very high

values of n. Note that while the percentage of points admitting a single discriminator ESS eventually decreases, the corresponding area where a mixture ESS exists seems to increase monotonically. In particular, this suggests that if the costs are less than half the benefits, our two ESS types cover the complete area in the limit as n grows large. However, even low values of n provide substantial chances for cooperation to be evolutionarily stable.

	$0 < r < 1$		$0 < r < 1/2$	
n	% single ESS	% ESS	% single ESS	% ESS
2	12,4	33,0	6,9	42,7
3	15,0	37,2	10,3	51,1
4	15,9	39,1	12,1	55,5
5	16,5	40,7	13,3	58,9
6	16,8	41,8	14,1	61,4
7	17,0	42,7	14,6	63,4
8	17,1	43,4	15,0	64,9
9	17,2	44,2	15,3	66,4
10	17,2	44,8	15,6	67,7
11	17,2	45,2	15,8	68,7
12	17,2	45,8	16,0	69,8
13	17,2	46,2	16,1	70,7
14	17,1	46,6	16,2	71,5
15	17,1	47,0	16,3	72,3
20	16,8	48,6	16,6	75,3
50	14,7	53,4	16,9	84,1
100	12,4	57,0	17,7	89,8
200	9,8	60,6	16,2	94,3
500	6,6	65,3	11,9	97,8
1000	4,5	68,7	8,5	98,8
5000	1,4	75,9	2,7	99,7

Table 1: Percentage of points (α, r) with existence of a single discriminator ESS or an ESS mixture of two neighboring discriminators (first column: single, second column: single or mixture)

4 Conclusions

Explanations of cooperation relying on the image scoring mechanism of indirect reciprocity have met with two points of criticism. The first was that within an *unrestricted* strategy space cooperation based on first-order information is never an ESS. This, of course, cannot be circumvented. E.g. if we introduce unresponsive strategies randomizing between C and D into our space of threshold strategies, then cooperation falls apart. Even if p_i^{fix} happens to be the payoff-maximizing cooperation rate, a static strategy cooperating with probability p_i^{fix} on each interaction constitutes a neutral mutant and can drift into the incumbent population, paving the way for unconditional defectors. Within the restricted strategy spaces studied in the literature, a second point of critique was that implementation errors destabilize cooperation based on discrimination. We have shown that this second point of criticism is not valid in general, applying only to the case of binary scoring. We have shown that if at least two observations are made, implementa-

tion errors may stabilize the evolution of cooperation via indirect reciprocity. This holds even in the presence of small costs of reputation transmission, i.e. under assumptions where higher-order assessment rules don't work.

An obvious limit of the present analysis is that it is a static one. The ESS property of a discriminator tells us nothing about the size of its basin of attraction under a learning or evolutionary dynamics. For the same reason we have to leave open the question what exactly happens when parameters are in a region where neither a homogeneous discriminator population nor a mixture of two neighboring discriminators are evolutionarily stable. Numerical simulations might shed further light on these questions.

References

[1] Alexander RD (1987) The Biology of Moral Systems. NewYork: Aldine deGruyter.

[2] Awaya Y (2014) Community enforcement with observation costs. J Econ Theory 154: 173–186.

[3] Axelrod R, Hamilton W (1981) The evolution of cooperation. Science 211: 1390-1396.

[4] Berger U (2009) Simple scaling of cooperation in donor-recipient games. BioSystems 97: 165-167.

[5] Berger U (2011) Learning to cooperate via indirect reciprocity. Games Econ Behav 72: 30-37.

[6] Binmore K (1992) Fun and Games. D. C. Heath, Lexington, MA.

[7] Bolton G, Katok E, Ockenfels A (2005) Cooperation among strangers with limited information about reputation. J Public Econ 89: 1457-1468.

[8] Boyd R (1989) Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. J Theor Biol 136: 47-56.

[9] Boyd R, Richerson P (1988) The evolution of reciprocity in sizable groups. J Theor Biol 132: 337-356.

[10] Brandt H, Sigmund K (2004) The logic of reprobation: Assessment and action rules for indirect reciprocation. J Theor Biol 231: 475-486.

[11] Brandt H, Sigmund K (2005) Indirect reciprocity, image scoring, and moral hazard. Proc Natl Acad Sci USA 102: 2666-2670.

[12] Broom M, Rychtar J (2013) Game-Theoretical Models in Biology. CRC Press.

[13] Cressman R (2003) Evolutionary Dynamics and Extensive Form Games. MIT Press.

[14] Ellison, G (1994) Cooperation in the Prisoner's Dilemma with anonymous random matching. Rev Econ Stud 61: 567-588.

[15] Engelmann D, Fischbacher U (2009) Indirect reciprocity and strategic reputation building in an experimental helping game. Games Econ Behav 67: 399-407.

[16] Fishman MA (2003) Indirect reciprocity among imperfect individuals. J Theor Biol 225: 285-292.

[17] Fudenberg D, Maskin E (1990) Evolution and cooperation in noisy repeated games. Am Econ Rev 80: 274-279.

[18] Gintis H (2000) Game Theory Evolving. Princeton University Press.

[19] Heller Y, Mohlin E (2015) Observations on cooperation. Working paper, University of Oxford.

[20] Hofbauer J, Sigmund K (1998) Evolutionary Games and Population Dynamics. Cambridge University Press.

[21] Kandori M (1992) Social norms and community enforcement. Rev Econ Stud 59: 63-80.

[22] Kurokawa S, Ihara Y (2009) Emergence of cooperation in public goods games. Proc Biol Sci 276: 1379-1384.

[23] Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. Proc Biol Sci 268: 745-753.

[24] Maynard Smith J (1982) Evolution and the Theory of Games. Cambridge University Press.

[25] Maynard Smith J, Price GR (1973) The logic of animal conflict. Nature 246: 15-18

[26] Milinski M, Semmann D, Bakker TCM, Krambeck HJ (2001) Cooperation through indirect reciprocity: Image scoring or standing strategy? Proc R Soc Lond B 268: 2495-2501.

[27] Nowak MA (2006) Five Rules for the Evolution of Cooperation. Science 314: 1560-1563.

[28] Nowak MA, Sigmund K (1998a) Evolution of indirect reciprocity by image scoring. Nature 393: 573-577.

[29] Nowak MA, Sigmund K (1998b) The dynamics of indirect reciprocity. J Theor Biol 194: 561-574.

[30] Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437: 1291-1298.

[31] Ohtsuki H (2004) Reactive strategies in indirect reciprocity. J Theor Biol 227: 299-314.

[32] Ohtsuki H, Iwasa Y (2004) How should we define goodness? Reputation dynamics in indirect reciprocity. J Theor Biol 231: 107-120.

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

[34] Okuno-Fujiwara M, Postlewaite A (1995) Social norms and random matching games. Games Econ Behav 9: 79-109.

[35] 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.

[36] Samuelson L (1998) Evolutionary Games and Equilibrium Selection. MIT Press.

[37] Sandholm W (2010) Population Games and Evolutionary Dynamics. MIT Press.

[38] Seinen I, Schram A (2006) Social status and group norms: Indirect reciprocity in a repeated helping experiment. European Econ Rev 50: 581-602.

[39] Sigmund K (2012) Moral assessment in indirect reciprocity. J Theor Biol 299: 25-30.

[40] Sugden R (1986) The Economics of Rights, Co-operation and Welfare. Basil Blackwell, Oxford.

[41] Suzuki S, Kimura H (2013) Indirect reciprocity is sensitive to costs of information transfer. Sci Rep 3, article no. 1435. doi:10.1038/srep01435

[42] Suzuki S, Akiyama E (2007) Three-person game facilitates indirect reciprocity under image scoring. J Theor Biol 249: 93–100.

[43] Suzuki S, Akiyama E (2008) Evolutionary stability of first-order-information indirect reciprocity in sizable groups. Theor Popul Biol 73:426-436.

[44] Takahashi S (2010) Community enforcement when players observe partners' past play. J Econ Theory 145: 42-62.

[45] Tanabe S, Suzuki H, Masuda N (2013) Indirect reciprocity with trinary reputations. J Theor Biol 317: 338–347.

[46] Taylor M (1976) Anarchy and Cooperation. John Wiley & Sons.

[47] Trivers RL (1985) Social Evolution. Menlo Park, CA: Benjamin Cummings.

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

[49] van Segbroeck S, Pacheco J, Lenaerts T, Santos F (2012) Emergence of fairness in repeated group interactions. Phys Rev Lett 108: 158104.

[50] van Veelen M (2012) Robustness against indirect invasions. Games Econ Behav 74, 382-393.

[51] Wedekind C, Milinski M (2000) Cooperation through image scoring in humans. Science 288: 850-852.

[52] Weibull J (1997) Evolutionary Game Theory. MIT press.

Appendix

Lemma 3 (Interval of cost-benefit ratios)**.** *Let* $r_i^{\min} = \mathrm{sl}_{f_i}(p_i^{\mathrm{fix}}, f_{i+1}(p_i^{\mathrm{fix}}))$ *and* $r_i^{\max} = \min\left(1, \mathrm{sl}_{f_i}(f_{i-1}(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}})\right)$*. Then*

$$r_i^{\min} < r < r_i^{\max} \implies i\text{-discr. is ESS} \implies r_i^{\min} \leq r \leq r_i^{\max}. \tag{12}$$

Proof. First we observe

$$\begin{aligned} \pi(m|i) - \pi(i|i) &\overset{(8)}{=} f_i(f_m(p_i^{\mathrm{fix}})) - r f_m(p_i^{\mathrm{fix}}) - f_i(f_i(p_i^{\mathrm{fix}})) + r f_i(p_i^{\mathrm{fix}}) \\ &= f_i(f_m(p_i^{\mathrm{fix}})) - f_i(p_i^{\mathrm{fix}}) \; - \; r\left(f_m(p_i^{\mathrm{fix}}) - p_i^{\mathrm{fix}}\right). \end{aligned}$$

Hence,

$$\begin{aligned} \pi(m|i) \leq \pi(i|i) \quad &\Leftrightarrow \quad f_i(f_m(p_i^{\mathrm{fix}})) - f_i(p_i^{\mathrm{fix}}) \leq r\left(f_m(p_i^{\mathrm{fix}}) - p_i^{\mathrm{fix}}\right) \\ &\Leftrightarrow \quad \begin{cases} \text{for } m > i: & r \geq \frac{f_i(f_m(p_i^{\mathrm{fix}})) - f_i(p_i^{\mathrm{fix}})}{f_m(p_i^{\mathrm{fix}}) - p_i^{\mathrm{fix}}} = \mathrm{sl}_{f_i}(p_i^{\mathrm{fix}}, f_m(p_i^{\mathrm{fix}})) \\ \text{for } m < i: & r \leq \frac{p_i^{\mathrm{fix}} - f_i(f_m(p_i^{\mathrm{fix}}))}{p_i^{\mathrm{fix}} - f_m(p_i^{\mathrm{fix}})} = \mathrm{sl}_{f_i}(f_m(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}}) \end{cases} \end{aligned}$$

and the analogous equivalence holds for the strict inequality. This proves

$$\tilde{r}_i^{\min} < r < \tilde{r}_i^{\max} \Longrightarrow i\text{-discr. is ESS} \Longrightarrow \tilde{r}_i^{\min} \leq r \leq \tilde{r}_i^{\max},$$

where

$$\tilde{r}_i^{\min} := \max_{m>i} \mathrm{sl}_{f_i}(p_i^{\mathrm{fix}}, f_m(p_i^{\mathrm{fix}})) \;\text{ and }\; \tilde{r}_i^{\max} := \min_{m<i} \mathrm{sl}_{f_i}(f_m(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}}).$$

It remains to be shown that $r_i^{\min} = \tilde{r}_i^{\min}$ and $r_i^{\max} = \tilde{r}_i^{\max}$.

For the first equality, $r_i^{\min} = \tilde{r}_i^{\min}$, we show that the slope $\mathrm{sl}_{f_i}(p_i^{\mathrm{fix}}, f_m(p_i^{\mathrm{fix}}))$ is decreasing in m for $m > i$. However, this follows from the observations in section 2.2. Since at the stable fixed point p_i^{fix} the function $f_i(.)$ intersects the diagonal $g(p) = p$ from above and $f_i(0) = 0$, p_i^{fix} must be in the concave part of $f_i(.)$, i.e.

$$p_i^{\mathrm{fix}} \geq \tilde{p}_i. \tag{13}$$

Hence, $f_i'(p)$ is non-increasing on the whole interval $[p_i^{\mathrm{fix}}, 1]$. This implies that the slope $\mathrm{sl}_{f_i}(p_i^{\mathrm{fix}}, p)$ is also non-increasing in p on $[p_i^{\mathrm{fix}}, 1]$ because it is the average of $f'(.)$ on $[p_i^{\mathrm{fix}}, p]$. The maximum slope is attained by the smallest p. This concludes the proof of this step because $f_m(p_i^{\mathrm{fix}})$ is by definition increasing in m.

For the second equality, $r_i^{\max} = \tilde{r}_i^{\max}$, we consider the function $h(p) := \mathrm{sl}_{f_i}(p, p_i^{\mathrm{fix}})$. First, we prove analogously to above that $h(p)$ is non-increasing on $[\tilde{p}_i, p_i^{\mathrm{fix}}]$ because $f_i'(.)$ is non-increasing on this interval.

$$h'(p) = \frac{-f_i'(p)(p_i^{\mathrm{fix}} - p) + \int_p^{p_i^{\mathrm{fix}}} f_i'(t)\,\mathrm{d}t}{(p_i^{\mathrm{fix}} - p)^2} \leq \frac{-f_i'(p)(p_i^{\mathrm{fix}} - p) + \int_p^{p_i^{\mathrm{fix}}} f_i'(p)\,\mathrm{d}t}{(p_i^{\mathrm{fix}} - p)^2} = 0$$

Let p_i^* be the intersection point of f_i and $g(p) = p$ such that for all p with $p_i^* < p < p_i^{\mathrm{fix}}$ we have $f_i(p) > p$. We want to prove that $h(p)$ is non-increasing on $[p_i^*, p_i^{\mathrm{fix}}]$. If $p_i^* > \tilde{p}_i$, we are done. Otherwise, we still have to prove the statement for the interval $[p_i^*, \tilde{p}_i]$. Note that we have $f_i'(p_i^*) \geq 1$ because $f_i(p_i^*) = p_i^*$ and $f_i(p_i^* + \varepsilon) > p_i^* + \varepsilon$ for small $\varepsilon > 0$. Because we are in the convex part of f_i, f_i' is increasing. Hence, $f'(p) \geq 1$ for every $p \in [p_i^*, \tilde{p}_i]$. This implies

$$\begin{aligned} h'(p) \quad &= \quad \frac{-f_i'(p)(p_i^{\mathrm{fix}} - p) + p_i^{\mathrm{fix}} - f_i(p)}{(p_i^{\mathrm{fix}} - p)^2} \\ &\overset{f_i(p) \geq p}{\leq} \quad \frac{-f_i'(p)(p_i^{\mathrm{fix}} - p) + p_i^{\mathrm{fix}} - p}{(p_i^{\mathrm{fix}} - p)^2} \quad = \quad \frac{1 - f_i'(p)}{p_i^{\mathrm{fix}} - p} \quad \overset{f_i'(p) \geq 1}{\leq} \quad 0. \end{aligned}$$

We have proved that $h(p)$ is non-increasing on $[p_i^*, p_i^{\mathrm{fix}}]$.

If $p_i^* > 0$, then $f_i(p) \leq p$ on $[0, p_i^*]$ because f_i intersects $g(p) = p$ from below in p_i^* and f_i is first convex on $[0, p_i^*]$, then possibly followed by a concave part. Hence,

$$\min_{m<i, f_m(p_i^{\mathrm{fix}}) \leq p_i^*} \mathrm{sl}_{f_i}(f_m(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}}) \geq 1.$$

Because $f_{-1}(p_i^{\mathrm{fix}}) = 0$, we even have the equality $\min_{m<i, f_m(p_i^{\mathrm{fix}}) \leq p_i^*} \mathrm{sl}_{f_i}(f_m(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}}) = 1$ and it holds also for $p_i^* = 0$. We can conclude

$$\begin{aligned} \tilde{r}_i^{\max} \quad &= \quad \min_{m<i} \mathrm{sl}_{f_i}(f_m(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}}) \quad = \quad \min\left(1, \min_{m<i, f_m(p_i^{\mathrm{fix}}) \geq p_i^*} \mathrm{sl}_{f_i}(f_m(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}})\right) \\ &= \quad \min\left(1, \mathrm{sl}_{f_i}(f_{i-1}(p_i^{\mathrm{fix}}), p_i^{\mathrm{fix}})\right) \quad = \quad r_i^{\max} \end{aligned}$$

Note that the last equality holds also if $f_{i-1}(p_i^{\mathrm{fix}}) < p_i^*$ where the minimum equals 1. This concludes the proof. □

Lemma 4 (Non-overlapping ESS-regions)**.** *For $i, j \in \{1, \dots, n-1\}$ and $i \neq j$, we have $R_i \cap R_j = \emptyset$. More precisely, $i < j$ and $0 < \alpha < \alpha_i^{\max}$ imply $r_j^{\max} < r_i^{\min}$.*

Proof. The proof uses the following observation about the fixed point p_i^{fix} and the inflection point $\tilde{p}_i = \frac{n-i-1}{n-1}$ as described around (7):

Lemma 5. *For $1 \leq i \leq n-1$, $p \in [\tilde{p}_i, 1)$ or $p \in [p_i^{\text{fix}}, 1)$ implies $f'_{i+1}(p) < f'_i(p)$.*

Proof of Lemma 5. For $i = n-1$ we have $f'_{i+1}(p) = f'_n(p) = 0 < f_i(p)$. Now, assume $1 \leq i \leq n-2$.

$$\begin{aligned} f'_i(p) - f'_{i+1}(p) &\overset{(5)}{=} (1-\alpha)(n-i)\frac{n!}{i!(n-i)!}p^{n-i-1}(1-p)^i \\ &\quad - (1-\alpha)(n-i-1)\frac{n!}{(i+1)!(n-i-1)!}p^{n-i-2}(1-p)^{i+1} \\ &= (1-\alpha)\frac{n!}{i!(n-i-2)!}p^{n-i-2}(1-p)^i\left(\frac{1}{n-i-1}p - \frac{1}{i+1}(1-p)\right) \end{aligned}$$

This shows

$$\begin{aligned} & f'_i(p) - f'_{i+1}(p) > 0 \quad \Leftrightarrow \quad \frac{1}{n-i-1}p - \frac{1}{i+1}(1-p) > 0 \\ \Leftrightarrow \quad & p\left(\frac{1}{n-i-1} + \frac{1}{i+1}\right) > \frac{1}{i+1} \quad \Leftrightarrow \quad p\left(\frac{i+1}{n-i-1} + 1\right) > 1 \\ \Leftrightarrow \quad & p > \frac{n-i-1}{n} \quad \Leftarrow \quad p \geq \tilde{p}_i \overset{(7)}{=} \frac{n-i-1}{n-1}. \end{aligned}$$

Within the proof of Lemma 3 we showed in (13) that $p_i^{\text{fix}} > \tilde{p}_i$. This proves that the implication holds for the precondition $p \geq p_i^{\text{fix}}$, too. □

To prove Lemma 4 it now suffices to prove the statement for $j = i+1$. Let $0 < \alpha < \alpha_i^{\max}$. The inequality $f_{i+1}(p) > f_i(p)$ for $0 < p < 1$, the monotonicity of both, f_i and f_{i+1}, $p_i^{\text{fix}} < p_{i+1}^{\text{fix}}$, and the fixed point properties of p_i^{fix} and p_{i+1}^{fix} imply

$$\begin{aligned} & p_i^{\text{fix}} = f_i(p_i^{\text{fix}}) \leq \quad f_{i+1}(p_i^{\text{fix}}) \quad \leq f_{i+1}(p_{i+1}^{\text{fix}}) = p_{i+1}^{\text{fix}} \\ \text{and} \qquad & p_i^{\text{fix}} = f_i(p_i^{\text{fix}}) \leq \quad f_i(p_{i+1}^{\text{fix}}) \quad \leq f_{i+1}(p_{i+1}^{\text{fix}}) = p_{i+1}^{\text{fix}}. \end{aligned} \tag{14}$$

Because of

$$p_i^{\text{fix}} \overset{(13)}{>} \tilde{p}_i \overset{(7)}{=} \frac{n-i-1}{n-1} > \frac{n-i-2}{n-1} \overset{(7)}{=} \tilde{p}_{i+1}$$

both cooperation functions f_i and f_{i+1} are concave on $[p_i^{\text{fix}}, 1]$. We have

$$\begin{aligned} r_i^{\min} &\overset{\text{Lemma 3}}{=} \mathrm{sl}_{f_i}(p_i^{\text{fix}}, f_{i+1}(p_i^{\text{fix}})) \overset{(14),\ f_i \text{ concave}}{\geq} \mathrm{sl}_{f_i}(p_i^{\text{fix}}, p_{i+1}^{\text{fix}}) \\ &\overset{\text{Lemma 5}}{>} \mathrm{sl}_{f_{i+1}}(p_i^{\text{fix}}, p_{i+1}^{\text{fix}}) \overset{(14),\ f_{i+1} \text{ concave}}{\geq} \mathrm{sl}_{f_{i+1}}(f_i(p_{i+1}^{\text{fix}}), p_{i+1}^{\text{fix}}) \\ &\geq \min\left(1, \mathrm{sl}_{f_{i+1}}(f_i(p_{i+1}^{\text{fix}}), p_{i+1}^{\text{fix}})\right) \overset{\text{Lemma 3}}{=} r_{i+1}^{\max} \end{aligned}$$

□

Proof of Theorem 2

Proof. Let q be the ratio of i-discriminators in population equilibrium. To simplify notation, we now denote the cooperation rates of the two discriminators in the steady state again by p_i and p_{i+1} instead of by $p_{i,i+1}^{\text{fix}}$ and $p_{i+1,i}^{\text{fix}}$. Moreover, we define

$$\tilde{f}_j(p_i, p_{i+1}) := q f_j(p_i) + (1-q) f_j(p_{i+1}) \tag{15}$$

for any $j \in \{-1, \ldots, n\}$.

In the steady state of cooperation rates, the right-hand sides of (10) must be zero, which is equivalent to

$$\tilde{f}_i(p_i, p_{i+1}) = p_i \text{ and } \tilde{f}_{i+1}(p_i, p_{i+1}) = p_{i+1}. \tag{16}$$

It will be useful to define the combined cooperation function of the whole population by

$$f_{i,i+1}(p) := q f_i(p) + (1-q) f_{i+1}(p). \tag{17}$$

In a mixed population equilibrium, the payoffs of both types of discriminators must be equal, since otherwise the discriminator with the higher payoff would increase in frequency. The payoff relation, again, has a useful slope formulation.

$$\begin{aligned} & \pi(i+1|i,i+1) = \pi(i|i,i+1) \qquad (18) \\ \Leftrightarrow\ & q f_i(p_{i+1}) + (1-q) f_{i+1}(p_{i+1}) - r p_{i+1} = q f_i(p_i) + (1-q) f_{i+1}(p_i) - r p_i \\ \overset{(17)}{\Leftrightarrow}\ & f_{i,i+1}(p_{i+1}) - r p_{i+1} = f_{i,i+1}(p_i) - r p_i \\ \Leftrightarrow\ & f_{i,i+1}(p_{i+1}) - f_{i,i+1}(p_i) = r(p_{i+1} - p_i) \\ \Leftrightarrow\ & \mathrm{sl}_{f_{i,i+1}}(p_i, p_{i+1}) = r. \end{aligned}$$

Note that the same equivalences hold, if we replace "=" by "<" or ">",

$$\pi(i+1|i,i+1) \gtrless \pi(i|i,i+1) \quad \Leftrightarrow \quad \mathrm{sl}_{f_{i,i+1}}(p_i, p_{i+1}) \gtrless r. \tag{19}$$

If the slope exceeds r, the $(i+1)$-discriminator has a payoff advantage and q decreases. This in turn increases both steady state cooperation rates p_i and p_{i+1}, since the right-hand sides of (10) are decreasing in q. As a consequence, the slope-term decreases,

since the cooperation rates are in the concave part of f_i and f_{i+1}, and hence of $f_{i,i+1}$. Analogous arguments show that the slope-term is increased, if it is below r. These arguments show that the mixed equilibrium population ratio q is unique.

The rest of the proof is very similar to the proof of Lemma 3. We want to prove the following equivalent statements[16] for $m \neq i$:

$$\begin{aligned}
& \pi(m|i,i+1) < \pi(i|i,i+1) = \pi(i+1|i,i+1) \qquad (20)\\
\Leftrightarrow\ & f_{i,i+1}(\tilde{f}_m(p_i,p_{i+1})) - r\tilde{f}_m(p_i,p_{i+1}) < f_{i,i+1}(p_i) - rp_i\\
\Leftrightarrow\ & f_{i,i+1}(\tilde{f}_m(p_i,p_{i+1})) - f_{i,i+1}(p_i) < r(\tilde{f}_m(p_i,p_{i+1}) - p_i)\\
\Leftrightarrow\ & \begin{cases} \mathrm{sl}_{f_{i,i+1}}(p_i,\tilde{f}_m(p_i,p_{i+1})) < r & \text{for } \tilde{f}_m(p_i,p_{i+1}) > p_i \\ \mathrm{sl}_{f_{i,i+1}}(\tilde{f}_m(p_i,p_{i+1}),p_i) > r & \text{for } \tilde{f}_m(p_i,p_{i+1}) < p_i \end{cases}\\
\Leftrightarrow\ & \begin{cases} \mathrm{sl}_{f_{i,i+1}}(p_i,\tilde{f}_m(p_i,p_{i+1})) < r & \text{for } m > i+1 \\ \mathrm{sl}_{f_{i,i+1}}(\tilde{f}_m(p_i,p_{i+1}),p_i) > r & \text{for } m < i \end{cases}
\end{aligned}$$

The last equivalence follows from the monotonicity of $f_j(p)$ in j and (16).

It is simple to prove that the statements in (20) hold for $m > i+1$. In the proof of Lemma 3 we have already used the inequality $p_i^{\mathrm{fix}} > \tilde{p}_i$ from (13), which means that $f_i(.)$ is concave on the whole interval $[p_i^{\mathrm{fix}}, 1]$. Since $\tilde{p}_{i+1} < \tilde{p}_i$, also $f_{i+1}(.)$ is concave on that interval[17]. Hence, $f_{i,i+1}(.)$ is concave, too. From $\tilde{f}_m(p_i,p_{i+1}) > \tilde{f}_{i+1}(p_i,p_{i+1}) = p_{i+1}$, we get

$$\mathrm{sl}_{f_{i,i+1}}(p_i,\tilde{f}_m(p_i,p_{i+1})) \ < \ \mathrm{sl}_{f_{i,i+1}}(p_i,p_{i+1}) \ \overset{(18)}{=} \ r.$$

For the case $m < i$, let us first exclude the special case $i = n-1$. Hence, since $0 \leq m < i$ we now consider $1 \leq i \leq n-2$. We can use the very same argumentation for $f_{i,i+1}$ which was used in the proof of Lemma 3 for f_i. In order to do so, we have to show that $f_{i,i+1}$ has the same crucial properties as f_i:

1. $f_{i,i+1}$ is strictly increasing from $f_{i,i+1}(0) = 0$ to $f_{i,i+1}(1) = 1-\alpha$.
2. $f_{i,i+1}$ cuts the line $g(p) = p$ from above at a point $p_{i,i+1}^{\mathrm{fix}}$.
3. There exists a fixed point of $f_{i,i+1}$ smaller than $p_{i,i+1}^{\mathrm{fix}}$. Let $p_{i,i+1}^{*}$ be the largest such fixed point.
4. $f_{i,i+1}$ is convex up to a value $\tilde{p}_{i,i+1} \in [0,1]$ and then concave.

1. holds because $f_{i,i+1}$ is a mixture of two functions which have this property. 2. holds for a value $p_{i,i+1}^{\mathrm{fix}} \in [p_i^{\mathrm{fix}}, p_{i+1}^{\mathrm{fix}}]$ because at p_i^{fix} we have $f_i(p_i^{\mathrm{fix}}) = p_i^{\mathrm{fix}}$ and $f_{i+1}(p_i^{\mathrm{fix}}) > p_i^{\mathrm{fix}}$, hence $f_{i,i+1}(p_i^{\mathrm{fix}}) > p_i^{\mathrm{fix}}$. Analogously, one can show that $f_{i,i+1}(p_{i+1}^{\mathrm{fix}}) < p_{i+1}^{\mathrm{fix}}$. Hence, the

[16] Here, $\pi(j|i,i+1)$ denotes the payoff of a single j-discriminator in the equilibrium population mixture of i- and $(i+1)$-discriminators.

[17] Note that since $m > i+1$ we know that $i+1 < n$.

fixed point $p^{\text{fix}}_{i,i+1}$ exists due to continuity of $f_{i,i+1}$. Furthermore, 4. will show that this fixed point is unique. 3. holds because at least $p = 0$ satisfies all conditions.

To prove 4., note that for $p < \tilde{p}_{i+1} < \tilde{p}_i$ we know that $f''_{i+1}(p) > 0$ and $f''_i(p) > 0$, and hence $f''_{i,i+1}(p) > 0$. Actually, since $q > 0$ we even know $f''_{i,i+1}(\tilde{p}_{i+1}) > 0$. Analogously, one can prove $f''_{i,i+1}(p) < 0$ for $p \in [\tilde{p}_i, 1]$. What happens between $\tilde{p}_{i+1}$ and $\tilde{p}_i$? One can use (6) to show that the sign of $f''_{i,i+1}(.)$ on $(0, 1)$ depends only on a function which is a quadratic polynomial in p. Hence, it can have at most two roots in $[\tilde{p}_{i+1}, \tilde{p}_i]$. However, the change of sign between $f''_{i,i+1}(\tilde{p}_{i+1})$ and $f''_{i,i+1}(\tilde{p}_i)$ shows that the number of roots must be odd. Hence, we have exactly one root in that interval which completes the proof of the 4th property.

Now, applying the analogous arguments as in the proof of Lemma 3, shows that each $p \leq p^*_{i,i+1}$ satisfies $\text{sl}_{f_{i,i+1}}(p, p_i) \geq 1 > r$, and each $p \in [p^*_{i,i+1}, p_i)$ satisfies $\text{sl}_{f_{i,i+1}}(p, p_i) > \text{sl}_{f_{i,i+1}}(p_i, p_{i+1}) = r$. In particular, this holds for $p := \tilde{f}_m(p_i, p_{i+1}) < \tilde{f}_i(p_i, p_{i+1}) = p_i$.

We still have to deal with the special case $i = n - 1$. Here, $f_{i,i+1}{=}f_{n-1,n}$ is a mixture of $f_{n-1}(p) = (1{-}\alpha)(1{-}(1{-}p)^n)$ which is concave everywhere and $f_n(p) = (1{-}\alpha)$. Since $q > 0$, the function $f_{i,i+1}$ is strictly increasing and concave on the whole interval $[0, 1]$. This proves that for every $p < p_i$ we have $\text{sl}_{f_{i,i+1}}(p, p_{i+1}) > \text{sl}_{f_{i,i+1}}(p_i, p_{i+1}) = r$, in particular for $p := \tilde{f}_m(p_i, p_{i+1}) < p_i$ like above. □