Department of Economics
Working Paper No. 168

# Evolutionary Stability of Indirect Reciprocity by Image Scoring

Ulrich Berger
Ansgar Grüne

February 2014

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

# Evolutionary Stability of Indirect Reciprocity by Image Scoring

Ulrich Berger[*] Ansgar Grüne[†]

February 18, 2014

**Abstract:** Indirect reciprocity describes a class of reputation-based mechanisms which may explain the prevalence of cooperation in groups where partners meet only once. The first model for which this has analytically been shown was the binary image scoring mechanism, where one's reputation is only based on one's last action. But this mechanism is known to fail if errors in implementation occur. It has thus been claimed that for indirect reciprocity to stabilize cooperation, reputation assessments must be of higher order, i.e. contingent not only on past actions, but also on the reputations of the targets of these actions. We show here that this need not be the case. A simple image scoring mechanism where more than just one past action is observed provides ample possibilities for stable cooperation to emerge even under substantial rates of implementation errors.

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

*JEL classification:* C72, D83

## 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

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

[†] Ansgar Grüne, Beethovenstr. 55, 53115 Bonn, Germany, ansgar.gruene@gmail.com

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 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, the reputation-based mechanism of *indirect reciprocity* (Trivers, 1971, Sugden, 1986, Alexander, 1987) is of primary interest.

Under indirect reciprocity, helping others enhances one's reputation, and help is primarily directed towards those with a high reputation. The costs of helping are more than offset by the benefits of being helped when in need, which aligns individual and social rationality of cooperation. Under strict rationality assumptions in a repeated-games framework with random matching this principle works via a process of community enforcement (Kandori, 1992). In a bounded-rationality framework, learning and evolutionary approaches have shown that a population of discriminators who base their decisions on their partner's reputation may successfully resist invasion attempts of defectors and unconditional cooperators. The first such approach to be formalized was Nowak and Sigmund's (1998a, 1998b) model of image scoring.

## 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. If only the last interaction of an individual is observed, the score becomes binary and discriminators assess other individuals as either *Good* or *Bad*, depending on whether or not they helped on their last interaction. In any interaction, discriminators then help those and only those which are assessed as Good. In updating an individual's reputation, the scoring rule relies only on the individual's behavior towards its last 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, but analytical results for the binary version of image scoring show that discriminators are only neutrally, but not evolutionarily stable (Nowak and Sigmund, 1998a,b). Indeed, Panchanathan and Boyd (2004) 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: an individual which refuses to help a "bad" opponent becomes "bad" itself.

## 1.3 Higher-order assessment rules

Panchanathan and Boyd (2004) 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). 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. 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). Later literature has largely focused on higher-order assessment rules, see the recent survey of Sigmund (2012).

In the last decade the overall picture has emerged that evolutionary stability of indirect reciprocity can only be established under higher-order assessment rules.[1] However, almost all of these higher-order assessment rules rely on the reputations of individuals being built and truthfully spread by word-of-mouth. If building or spreading a reputation is only slightly costly, then this results in another social dilemma and renders cooperation impossible, as shown by Suzuki and Kimura (2013).

All in all, the situation seems puzzling: Higher-order assessment rules are cognitively highly demanding and rely on costless and truthful reputation building, which makes it difficult for them to explain indirect reciprocity. The first-order assessment rule of image scoring, on the other hand, is theoretically unstable in general models. But indirectly reciprocal behavior in humans 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 From binary to multi-valued scores

Our answer in this paper is that the alleged instability of cooperation under image scoring is an artefact of the assumption of binary scores in the received analytical approaches to image scoring. We show that in an image scoring model with multi-valued scores, cooperation is indeed evolutionarily stable under a wide range of parameter values.

The dynamics of cooperation under image scoring have previously been studied by Berger (2011) for the special case of two observations of opponents' past actions and a low error

[1] There are 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), growing social networks (Brandt and Sigmund, 2005), interactions in larger groups (Suzuki and Akiyama, 2007, 2008), or trinary reputation values (Tanabe et al., 2013).

rate. However, this analysis was restricted to a three-strategy setting where the most tolerant of the discriminating strategies competed with unconditional cooperators and defectors only. Since the question of evolutionary stability heavily depends on the set of feasible strategies, a convincing analysis requires inclusion of all *threshold strategies*, i.e. strategies which cooperate if and only if the number of the opponent's defections in a sample of $n$ of his past actions does not exceed a certain threshold $i$. These threshold strategies include unconditional defectors (for $i = -1$) and cooperators (for $i = n$). This was also the universe of strategies studied in the original image scoring model of Nowak and Sigmund (1998a). Here we study evolutionary stability in settings with fixed but arbitrary observation sample size $n \geq 1$, including all $n + 2$ associated threshold strategies and also allowing for implementation errors.

# 2 Model

## 2.1 The donation game, errors, and threshold strategies

Consider a large population of individuals. Time $t$ is continuous and individuals are repeatedly and randomly matched in pairs to interact in the 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.[2] 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 donor intends to defect.

Before a donor implements his action, he is informed of his partner's choices in a random sample[3] of $n \geq 1$ past interactions where this partner was in the donor-role. 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. 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 ($i = -1$).

[2] This means that on each interaction, individuals play a Prisoner's Dilemma game with "equal gains from switching".

[3] For technical simplicity we assume sampling with replacement. While this makes it possible that some past action is sampled two or more times, it doesn't change the results.

### 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

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

Here, $F(i; n, 1-p)$ 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.

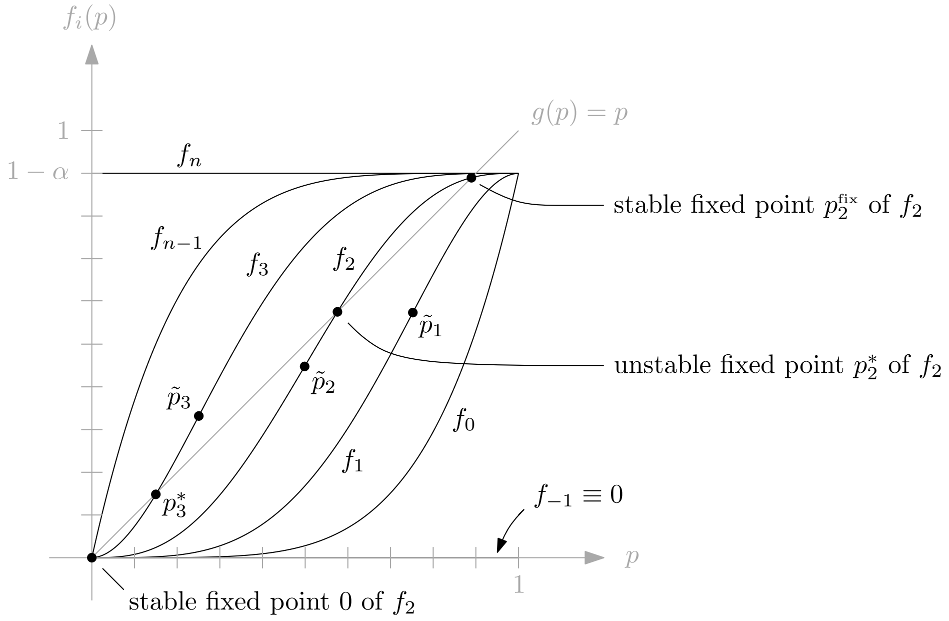


Figure 1: Cooperation functions $f_i(p)$ for $n = 5$ and $\alpha = 0.1$.

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

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

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

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

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 $i \in \{0, \dots, n-1\}$ 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 $i \in \{0, \dots, n-1\}$, $f_i'(.)$ 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 Fixed points

Consider now, for $i \in \{0, \dots, n-1\}$, a homogeneous population of $i$-discriminators. Assume that at time $t$ the past cooperation rate in the population is $p(t)$. As long as $f_i(p(t)) < p(t)$, this cooperation rate will decrease, and as long as $f_i(p(t)) > p(t)$, the cooperation rate will increase. Thus the overall cooperation rate will monotonically converge to a fixed point of the cooperation function $f_i$.

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 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 $\alpha$ is large enough. But if $\alpha$ 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 $i$-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 $\alpha$, $n \geq 2$, 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.[4] 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.

If $\alpha$ is small enough, self-cooperation is always obtained for the cases $i = n$ (an ALLC-population) and $i = n-1$. 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 newly born 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

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

always end up with a cooperation rate at the high-cooperation fixed point $p_i^{\text{fix}}$.

## 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 entry the incumbent's cooperation rate has already stabilized at $p_i^{\text{fix}}$.[5] We denote the payoff 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 $\pi(j|i) := b^{-1}\hat{\pi}(j|i)$.[6]

When a mutant $m$-discriminator enters an incumbent population of $i$-discriminators, the incumbents' overall cooperation rate remains at $p_i^{\text{fix}}$, 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

[5] This basically means that interactions take place on a much faster time scale than strategy adjustments.

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

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 can not be overcome.[7] 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 $\alpha$ and $1 \leq i \leq n-1$ are such that the $i$-discriminator is self-cooperative with cooperation rate $p_i^{\text{fix}}$. Self-cooperativeness implies that at $p_i^{\text{fix}}$ the 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 $p_i^{\text{fix}}$. 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 $\alpha > 0$ and $n \geq 2$. 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.*

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 $\alpha$ 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 $\alpha > 0$.

# 3 ESS Regions

## 3.1 Overview

The exact shape of the ESS-regions $R_i$ in the interior of the $\alpha$-$r$-square where an $i$-discriminator is an ESS, can be determined numerically from inequality (9). It turns out that for small $\alpha$ the open intervals of $r$-values guaranteeing the ESS-property for the $i$-discriminators can be extremely small. However, a sizable fraction of the $\alpha$-$r$-square

[7] 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.

consists of parameter combinations where some discriminator is an ESS. For $n = 5$ these ESS regions are depicted in Figure 2 as the green “leaves” originating from $(0, 0)$.

Note that relatively large values of $\alpha < 1$ can not readily be interpreted as probabilities of implementation errors of intended donations. Rather, high values of $\alpha$ indicate that individuals intending to help often simply lack the resources to do so. This suggests an interpretation of such a high-$\alpha$ population as a poor society. For very large values of $\alpha$, 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 $\alpha$-values arbitrarily close to 0.8 (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.

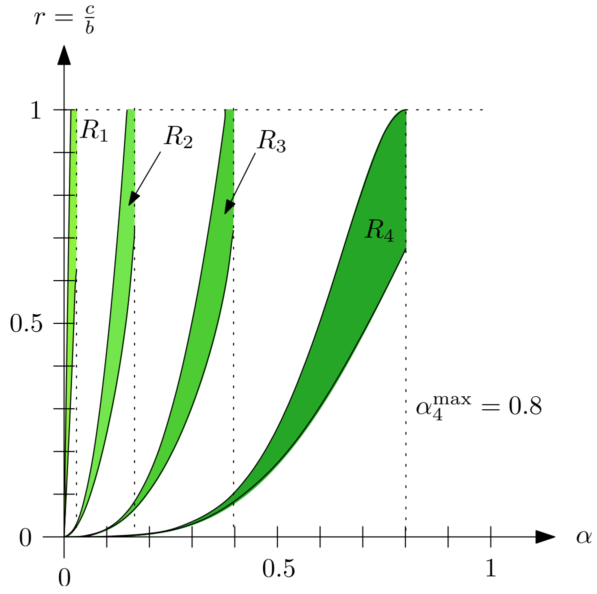


Figure 2: ESS regions $R_1, \dots, R_4$ for $n = 5$.

## 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 $\alpha$ is small enough. Increasing $\alpha$ scales down the cooperation functions in Figure 1 until the unstable fixed point $p_i^*$ and the stable fixed point $p_i^{\text{fix}}$ coincide and the diagonal is tangential to the cooperation function at this value. The value of $\alpha$ 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 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 $\alpha$. 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 Evolutionary 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 $(\alpha, r)$ such that there exists an $i \in \{1, \dots, n-1\}$ 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 evolutionary 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 evolutionary 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 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 overall cooperation rate of the $i$-discriminators will thus be moved into this direction. The analogous applies to the $(i+1)$-discriminator. In a large population this movement of cooperation rates can be approximated 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. [8]

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

### 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 $p^{\text{fix}}_{i,i+1}$ and $p^{\text{fix}}_{i+1,i}$, respectively. Therefore, the payoffs of an $i$-discriminator and an $(i+1)$-discriminator are given by

$$\begin{aligned} \pi_i &= q f_i(p^{\text{fix}}_{i,i+1}) + (1-q) f_{i+1}(p^{\text{fix}}_{i,i+1}) - r p^{\text{fix}}_{i,i+1}, \\ \pi_{i+1} &= q f_i(p^{\text{fix}}_{i+1,i}) + (1-q) f_{i+1}(p^{\text{fix}}_{i+1,i}) - r p^{\text{fix}}_{i+1,i}, \end{aligned} \tag{11}$$

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

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

The two payoffs can then be written as $\pi_i = f^q_{i,i+1}(p^{\text{fix}}_{i,i+1}) - r p^{\text{fix}}_{i,i+1}$ and $\pi_{i+1} = f^q_{i,i+1}(p^{\text{fix}}_{i+1,i}) - r p^{\text{fix}}_{i+1,i}$. Division by $p^{\text{fix}}_{i+1,i} - p^{\text{fix}}_{i,i+1}$ 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 $(p^{\text{fix}}_{i+1,i}, f^q_{i,i+1}(p^{\text{fix}}_{i+1,i}))$ and $(p^{\text{fix}}_{i,i+1}, f^q_{i,i+1}(p^{\text{fix}}_{i,i+1}))$, i.e. the difference $r - \mathrm{sl}_{f^q_{i,i+1}}(p^{\text{fix}}_{i,i+1}, p^{\text{fix}}_{i+1,i})$. In particular, equality of payoffs implies $\mathrm{sl}_{f^q_{i,i+1}}(p^{\text{fix}}_{i,i+1}, p^{\text{fix}}_{i+1,i}) = r$.

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

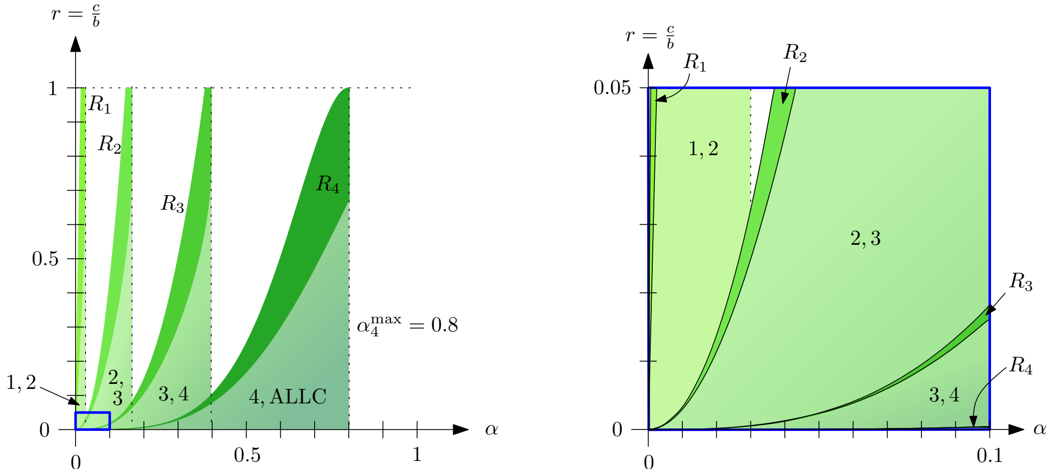


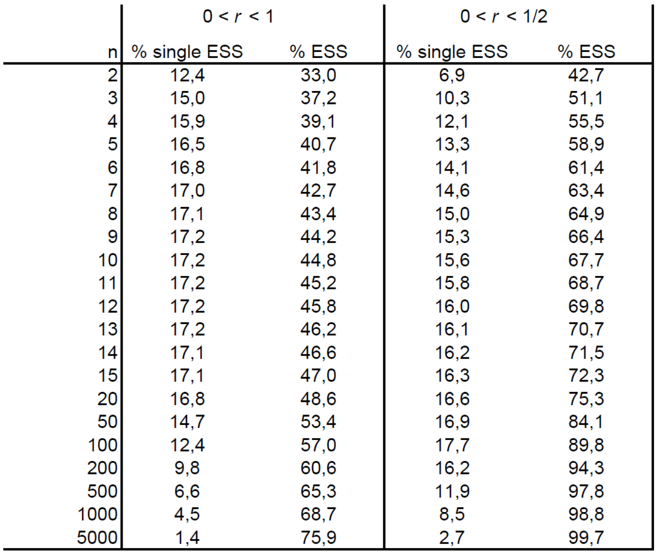
Figure 3: ESS regions of single discriminating strategies and mixtures of two discriminating strategies for $n = 5$. On the right side a zoom of the area close to the origin.

### 3.4 The chances for evolutionary stability

Given randomly selected values for $\alpha$ 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 evolutionary stable.

## 4 Conclusions

While image scoring was the very first explicit model of indirect reciprocity, the general version had hitherto only been studied numerically. Analytical results were available only for the binary version, which, however, is rather restricted. Discrimination always fails to sustain cooperation in the binary image scoring model with implementation er-



| | 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 $(\alpha, r)$ with existence of a single discriminator ESS or an ESS mixture of two neighboring discriminators (first column: single, second column: single or mixture)

rors. However, as we have shown in a simple model of the general version of image scoring, under small error rates every proper discriminating strategy except the most intolerant one is evolutionarily stable for some interval of cost-benefit ratios. Interestingly, cooperation can even be upheld by very tolerant discriminators in poor societies, i.e. under substantial rates of failed intended donations. A 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] Berger U (2009) Simple scaling of cooperation in donor-recipient games. BioSystems 97: 165-167.

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

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

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

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

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

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

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

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

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

[12] 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.

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

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

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

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

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

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

[19] Ohtsuki H, Iwasa Y (2006) The leading eight: Social norms that can maintain

cooperation by indirect reciprocity. J Theor Biol 239: 435-444.

[20] 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.

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

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

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

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

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

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

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

[28] Trivers RL (1971) Evolution of reciprocal altruism. Quarterly Rev Biol 46: 35-57.

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

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

## 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*

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

*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,

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

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

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

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 $\mathrm{sl}_{f_i}(p_i^{\text{fix}}, f_m(p_i^{\text{fix}}))$ is decreasing in $m$ for $m > i$. However, this follows from the observations in Section 2.2. Since at the stable fixed point $p_i^{\text{fix}}$ the function $f_i(.)$ intersects the diagonal $g(p) = p$ from above and $f_i(0) = 0$, $p_i^{\text{fix}}$ must be in the concave part of $f_i(.)$, i.e.

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

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

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

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

Let $p_i^*$ be the intersection point of $f_i$ and $g(p) = p$ such that for all $p$ with $p_i^* < p < p_i^{\text{fix}}$ we have $f_i(p) > p$. We want to prove that $h(p)$ is non-increasing on $[p_i^*, p_i^{\text{fix}}]$. 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

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

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

If $p_i^* > 0$, then $f_i(p) \le p$ 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,

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

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

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

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

**Lemma 4** (Non-overlapping ESS-regions)**.** *For $i, j \in \{1, \ldots, n-1\}$ 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 $p_i^{\text{fix}}$ and the inflection point $\tilde{p}_i = \frac{n-i-1}{n-1}$ as described around (7):

**Lemma 5.** *For $1 \le i \le n-1$, $p \in [\tilde{p}_i, 1)$ 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 $1 \le i \le n-2$.

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

This shows

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

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 $p \ge p_i^{\text{fix}}$, 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 $p_i^{\text{fix}}$ and $p_{i+1}^{\text{fix}}$ imply

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

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

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

□

## 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

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

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

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

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

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

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.

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

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

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

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[9] for $m \neq i$:

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

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 $p_i^{\text{fix}} > \tilde{p}_i$ from (13), which means that $f_i(.)$ is concave on the whole interval $[p_i^{\text{fix}}, 1]$. Since $\tilde{p}_{i+1} < \tilde{p}_i$, also $f_{i+1}(.)$ is concave on that interval[10]. 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 $p_{i,i+1}^{\text{fix}}$.

[9] 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.

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

3. There exists a fixed point of $f_{i,i+1}$ smaller than $p^{\text{fix}}_{i,i+1}$. 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 $p^{\text{fix}}_{i,i+1} \in [p^{\text{fix}}_i, p^{\text{fix}}_{i+1}]$ because at $p^{\text{fix}}_i$ we have $f_i(p^{\text{fix}}_i) = p^{\text{fix}}_i$ and $f_{i+1}(p^{\text{fix}}_i) > p^{\text{fix}}_i$, hence $f_{i,i+1}(p^{\text{fix}}_i) > p^{\text{fix}}_i$. Analogously, one can show that $f_{i,i+1}(p^{\text{fix}}_{i+1}) < p^{\text{fix}}_{i+1}$. Hence, the 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. □