Social Norms and the Indirect Evolution of Conditional Cooperation

Christian Traxler* and Mathias Spichtig†

September 28, 2010

Abstract

We develop a model of social norms and cooperation in large societies. Within this framework we use an indirect evolutionary approach to study the endogenous formation of preferences and the co-evolution of norm compliance. The multiplicity of equilibria that emerges in the presence of social norms, is linked to the evolutionary analysis: individuals face situations where many others cooperate and situations where cooperation fails. The evolutionary adaptation to such heterogeneous environments favors conditional cooperators, who condition their pro-social behavior on others' cooperation. As conditional cooperators respond flexibly to their environment, they dominate free-riders and unconditional cooperators.

Keywords: Conditional cooperation; indirect evolution; social norms; heterogeneous environments.

JEL classification: C70; Z13

* *Corresponding Author.* Max Planck Institute for Research on Collective Goods. Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. Phone/Fax: +49 (0)228 91416-69/-62; E-mail: traxler@coll.mpg.de

† Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam. Kruislaan 320, 1098 SM Amsterdam, The Netherlands. E-mail: spichtig@science.uva.nl

1 Introduction

Starting with Keser and van Winden (2000) and Fischbacher et al. (2001), economic research has pointed out the role of conditional cooperation in human behavior. People who follow this behavioral pattern condition their cooperation on the cooperativeness of others or on their beliefs about the behavior of others – they *"are willing to contribute the more to a public good, the more others contribute"* (Fischbacher et al., 2001, p. 397). There is now a solid body of empirical evidence which documents the prevalence of conditional cooperation (Gächter, 2007). Motivated by this evidence, several social preference models have been proposed (surveyed in Fehr and Schmidt, 2006) that are capable explaining of conditionally cooperative behavior. The question under what circumstances evolution fosters preferences, which then induce conditional cooperation, has gained little attention. The focus of this paper is to address exactly this question.

One possible way to capture conditional cooperation is based upon *social norms*.[1] Social norms are rules of conduct which are enforced by internal or external sanctions (Coleman, 1990). As the sanctions for a norm deviation are harsher the more people adhere to the norm (Traxler and Winter, 2009), a social norm for cooperation can trigger conditionally cooperative behavior. The present analysis incorporates such a concept of social norms into a model of voluntary public good provision in a large society. Within this framework we study the evolution of a cooperation norm and the coevolution of behavior. This allows us to discuss the prerequisites for the emergence of conditional cooperation. Our analysis thereby provides several novel elements.

First of all, the strength of the social norm reflected in the impact of norm-enforcing sanctions depends on the level of norm compliance in the society as well as on an individual specific level of norm sensitivity: some agents suffer more from sanctions than others do. For a given distribution of norm sensitivity in the population, the equilibrium level of cooperation derives endogenously. Similar to other models of social norms, there is scope for a multiplicity of equilibria: society could either coordinate on equilibrium states with a strong social norm and far-reaching cooperation or on states with weak norm-enforcement and widespread free-riding.

In a next step, we study the evolution of the norm. So far, the literature has mainly focused on actual behavior as the determinant of an endogenous norm strength (Akerlof, 1980; Lindbeck et al., 1999). In addition to this channel, we also consider an individual norm sensitivity as evolving endogenously. We model the evolution of the norm sensitiv-

[1] Other theoretical approaches which account for conditional cooperation are theories of conformity, inequity aversion and reciprocity, surveyed in Fehr and Schmidt (2006).

ity as an indirect evolutionary process.[2] Heterogeneous levels of norm sensitivity induce different behavioral patterns, which are associated with different levels of evolutionary success. Next to the economic payoff from free-riding and cooperation, the success is also determined by the norm-based costs for violating a norm. One might think of social disapproval as one mechanism behind these evolutionary costs. Depending on whether the level of disapproval in the society is sufficiently strong to outbalance the cost of cooperation, either the pro-social or the selfish behavior dominates in terms of evolutionary success. Accordingly, either higher or lower levels of norm sensitivity are evolutionarily more successful and spread within the society – e.g., by the vertical transmission of the social norm (socialization and education). In this vein, the distribution of the norm sensitivity evolves endogenously. Individual behavior, the level of cooperation within the population and the associated strength of sanctions (disapproval) evolves indirectly, along with the endogenous change in preferences. In an evolutionary equilibrium, the outcome is shaped by preferences and – at the same time – the outcome shapes these preferences.

We first discuss the evolutionary transmission of norms within a *homogenous environment*, associated with one particular equilibrium state of the public good game. There exists an evolutionary equilibrium with a distribution of norm-sensitivities such that free-riders and cooperators coexist. However, this equilibrium is unstable. Evolution will typically induce a decline in the norm sensitivity and cooperation will break down. In the evolutionary equilibrium the norm has eroded and nobody contributes to the public good.

This result changes once we incorporate the multiplicity of equilibria into the analysis. We focus on the case of a *heterogeneous environment*, in the sense that the population faces an equilibrium state with strong norm-compliance and a state with widespread norm violations, where both states are supported by one given distribution of preferences. Agents then interact in 'cooperative' and 'non-cooperative' situations, with a strong impact of sanctions in the former and a weak norm in the latter environment. One can think of many real-life situations which can be described as a heterogeneous social environment: people walk through clean and littered public parks (and may stick to an anti-littering norm), through nice and run-down neighborhoods (and are tempted to commit a crime, see Funk, 2005, Glaeser et al., 1996); we visit parties where nobody smokes, but also face some where people do smoke (Nyborg and Rege, 2003); we are confronted with charity projects, some of which receive more and others fewer donations (Frey and Meier, 2004); we sometimes give large tips and sometimes we completely avoid tipping (Azar, 2005); we work in firms where many co-workers cheat but we are also engaged in projects where others' exert high efforts (Ichino and Maggi, 2000).

[2]The indirect evolutionary approach was pioneered by Güth and Yaari (1992) and Güth (1995).

In a stylized model of such heterogeneous environments we observe three different types of behavior: free-riders, who violate the norm in both situations, unconditional cooperators, who always comply with the social norm, and conditional cooperators. These agents cooperate in the 'good' state, where many others follow the norm, but defect in the 'bad' state, where a majority free-rides. In the environment with a strong social norm, conditional cooperators avoid harsh sanctions, making them more successful than free-riders. In the environment where the norm is weak they free-ride and earn a higher evolutionary payoff than unconditional cooperators. Hence, the conditional strategy dominates both unconditional strategies in terms of evolutionary success. Norm transmission will favor conditional cooperators, since they react flexibly to their social environment. We characterize conditions under which this dominance of conditional cooperation forms a stable evolutionary equilibrium.

Rather than explaining the emergence of pro-social norms, our paper studies evolutionary forces that shape conditional cooperation. While there are several approaches to explain the origin of social norms and pro-social behavior (e.g., Corneo and Jeanne, 1997, Fershtman and Weiss, 1998), only Mengel (2008) discusses conditional cooperation in a similar context to ours. Her paper studies the impact of migration on an internalized norm for cooperation. For some degrees of population viscosity – which can be neatly linked to the level of integration in a society – she finds a stable evolutionary equilibrium, where norm-sensitive and norm-insensitive agents coexist. As in our study, norm-sensitive individuals behave conditionally cooperative: they start to defect if norm-insensitive agents become more frequent in the population. This protects conditional cooperators from getting exploited and supports their evolutionary success. The result and its intuition is similar to our findings for the case of heterogeneous environments. In Mengel's analysis, conditional cooperation is a response to the heterogeneity in selfish or norm-guided interaction partners. In our model, it is the heterogeneity in social environments related to different equilibrium states which supports the conditional behavior. This structural similarity in the results suggests that the role of heterogeneous environments as a driving force in the evolution of conditional cooperation provides a robust finding which generalizes to different frameworks.

Finally, our paper also contributes to the literature by introducing a technique from quantitative genetics which – to the best of our knowledge – is novel in evolutionary economics. The method, originally developed in Lande (1976), provides a simple tool to analyze the evolution of a *continuously* distributed trait – in our case, the norm sensitivity. We discuss the crucial assumptions of Lande's approach and show that our main findings are qualitatively robust to the application of standard replicator dynamics (Weibull, 1995). The fact that we study the evolution of a continuous distribution of preferences, instead

of a discrete number of types, also distinguishes our model from Mengel (2008) and other contributions in the field.

The remaining paper is structured as follows. We first study a model of social norms and cooperation in a large population. In Section 3, we introduce an evolutionary approach from quantitative genetics. We then apply this method on our model and discuss the evolution of social norms and cooperative behavior in a homogenous and in a heterogeneous environment. Section 5 discusses our findings and Section 6 concludes.

2 Social Norms and Cooperation

Consider a large society represented by a continuum of individuals $[0,1]$. Each agent i chooses $x^i \in \{0,1\}$, to contribute to the public good ($x^i = 1$, 'cooperate') or not to contribute ($x^i = 0$, 'free-ride'). The payoff $y(x^i)$ for strategy x^i is given by

$$y(x^i) = -x^i c \tag{1}$$

where $c > 0$ denotes the costs of the public good contribution. The action x^i additionally determines a payoff $z(x^i, n)$, where n denotes the share of free-riders in the society. This payoff is defined as

$$z(x^i, n) = \left(x^i - 1\right) s(n) \tag{2}$$

where $s(n)$ relates to the sanctions (or the withdrawal of rewards) an agent incurs if she violates the social norm. In principle, the origin of these sanctions could be internal, external or a mixture of both (Coleman, 1990). Within our framework, one might best think of the sanctions as deriving from internalized social norms. If an agent has internalized a cooperation norm, free-riding would be associated with emotions like guilt, remorse or the loss of self-esteem (Elster, 1998). Alternatively, one could interpret sanctions also as being external, e.g., deriving from social disapproval (Traxler and Winter, 2009).[3] Throughout our analysis we employ the following assumption:

Assumption A1: The function $s(n) : [0,1] \to \mathbb{R}_+$ is continuously differentiable on $n \in [0,1]$ with $s'(n) \leq 0$, $s(0) > 0$ and $s(1) = 0$.

Allowing the sanctions to depend on other agents' behavior captures the idea that the degree of norm compliance (co)determines the strength of the social norm and thereby the strength of norm-enforcement. Motivated by the evidence in Traxler and Winter (2009),

[3]Note that the present paper studies neither the origin of social norms nor sanctioning mechanisms. As explained above, we want to analyze conditions that – for exogenously given patterns of norm enforcement – support the evolution of conditional cooperation.

we follow the literature (Lindbeck et al., 1999, Mengel, 2008) and assume $s(n)$ to be non-increasing in n.[4] A deviant agent is supposed to suffer from weaker sanctions, as free-riding becomes more widespread: one feels less guilty about violating a norm, the more others do the same. The equivalent is supposed to hold for external sanctions. For the case of perfect norm compliance ($n = 0$), sanctions are strictly positive. In a society where everybody free-rides, however, the social norm has completely eroded. The moral connotation of 'wrong' (free-riding) and 'right' (contributing) – and so the sanctions for free-riders – have vanished.

2.1 Preferences

Let the preferences of agent i, defined over $y(.)$, $z(.)$ and the public good payoff $v(.)$, be given by an additive separable utility function

$$u^i(x^i, n) = y(x^i) + \theta^i z(x^i, n) + v(n), \tag{3}$$

with the individual specific parameter $\theta^i \in \mathbb{R}$ and $v' < 0$. We can interpret the parameter θ^i as the degree of internalization or norm sensitivity. While an agent with $\theta^i = 0$ is solely driven by the material payoff from the game, those with $\theta^i > 0$ also consider sanctions in their decisions.[5]

In a large population, a single decision maker takes n as given. Hence, agent i will cooperate iff $u^i(1, n) > u^i(0, n)$ which is equivalent to $\theta^i s(n) > c$. An individual contributes to the public good if the utility loss from the sanction dominates the costs of cooperation. This implies the threshold

$$\hat{\theta}(n) \equiv \frac{c}{s(n)}, \tag{4}$$

which divides society into norm-adhering and norm-breaking individuals. Those with $\theta^i > \hat{\theta}(n)$ cooperate, while those with $\theta^i \leq \hat{\theta}(n)$ free-ride.[6] The action x^i is then determined by an individual's norm sensitivity θ^i and the share of free-riders n,

$$x^i = x(\theta^i, n) = \begin{cases} 0 & \text{for } \theta^i \leq \hat{\theta}(n) \\ 1 & \text{for } \theta^i > \hat{\theta}(n) \end{cases} \tag{5}$$

[4] For a micro-foundation of this pattern, see the signaling game in Corneo (1997) and the analysis in Benabou and Tirole (2006).

[5] Agents with $\theta^i < 0$ hold anti-social preferences, as they derive benefits from a norm-violation. As will become clear in the following, we only include this latter group for technical convenience. Excluding negative values of θ would not change any of our results.

[6] The assumption that agents with $\theta^i = \hat{\theta}(n)$ will free-ride is not crucial for any of our results.

Note that the threshold $\hat{\theta}(n)$ is non-decreasing in n,

$$\frac{\partial \hat{\theta}(n)}{\partial n} \geq 0, \tag{6}$$

since $s'(n) \leq 0$. As more agents deviate from the norm, the sanctions associated with a norm violation become smaller. Hence, an agent who cooperates for low levels of n may turn into a free-rider for higher levels of n.[7] Those with $\theta^i \in (\hat{\theta}(0), \hat{\theta}(1))$ condition their cooperation on the behavior of others. They act as *conditional cooperators*. Agents with $\theta^i \leq \hat{\theta}(0)$, however, would always free-ride, irrespectively of other subjects' behavior. Allowing for a heterogeneity in θ, the model therefore captures the two main patterns of behavior typically found in experimental studies (Fischbacher et al. 2001).

2.2 Equilibrium

Let the cumulative distribution function of the parameter θ be given by $\Phi(\theta)$. The corresponding density function $\phi(\theta)$ has full support.[8]

Assumption A2: (i) The inverse function of the cumulative distribution is given by $\Phi^{-1}(n)$ for $n \in [0,1]$, with $\Phi^{-1}(n) \to -\infty$ for $n \to 0$ and $\Phi^{-1}(n) \to c/s(n)$ for $n \to 1$. (ii) $\exists\ n' \in (0,1) : \Phi^{-1}(n') > \hat{\theta}(n')$.

A social equilibrium state in such a society is given by a share of free-riders n^*, characterized by the fixed point equation

$$n^* = \Phi(\hat{\theta}(n^*)). \tag{7}$$

Lemma 1 *For any $s(n)$ and $\Phi(\theta)$ satisfying A1 and A2(i) there always exists an equilibrium with $n^* = 1$. If A2(ii) holds, there always exists at least one further equilibrium with $0 < n^* < 1$.*

Proof. See Appendix B.

An equilibrium constitutes a self-supporting share of norm-violators – the threshold $\hat{\theta}(n^*)$ is such that the share of agents with $\theta^i \leq \hat{\theta}(n^*)$ is exactly n^*. There always exists one equilibrium where nobody contributes, $n^* = 1$. The cooperation norm has eroded and everybody free-rides. Given that assumption A2(ii) holds, the strength of the norm sensitivity is distributed such that there exists a level of free-riding $0 < n < 1$, where the

[7] Assumption A1 implies $\hat{\theta}(n) \to \infty$ for $n \to 1$. Thus, there always exists a level of free-riding at which cooperators turn into free-riders. Hence, universally unconditional cooperators cannot be present here.

[8] Assumption A2(i) assures technical properties of $\Phi(\theta)$ that allow for an equilibrium and are in line with the distribution considered in Section 4; A2(ii) is discussed after Lemma 1.

maximum level of norm sensitivity among free-riders, $\Phi^{-1}(n)$, is above the cooperation threshold $\hat{\theta}(n)$. In this case, the system is characterized by a multiplicity of equilibria. In addition to the equilibrium with $n^* = 1$, there is at least one equilibrium with a positive share of contributors. A graphical representation of two possible scenarios is provided in Figure 1. While assumption A2(ii) is fulfilled for the example depicted in panel (a) of the Figure, it does not hold for the example in panel (b). In the first case, there are multiple equilibria, in the latter there is a unique equilibrium at $n^* = 1$.

Figure 1 about here

If the distribution $\Phi(\theta)$ is common knowledge, society coordinates on one of the possible equilibria. Alternatively one could consider $\Phi(\theta)$ to be unknown, but assume that agents can infer the behavior of other members in society from the public good level. Agents could then learn about the share of free-riders. As long as players base their decision on this share, society would converge into an asymptotically stable equilibrium, characterized by

$$\frac{\partial \Phi^{-1}(n^*)}{\partial n} \geq \frac{\partial \hat{\theta}(n^*)}{\partial n}. \tag{8}$$

In the following we call an equilibrium n^* an *a-stable* equilibrium state, if (8) holds for n^*. In the scenario depicted in panel (a) in Figure 1, there are two unstable (the one with n_c^* and another one at $n^* = 1$) and two a-stable equilibrium states: one with a low level of free-riding n_a^* and another one where free-riding is widespread, n_b^*. In panel (b) the only equilibrium, $n^* = 1$, is also stable, since the cumulative distribution approaches the $\hat{\theta}(n)$-curve 'from below' (thus condition (8) holds).

3 Evolutionary Quantitative Genetics

In the following, we will study the evolution of the distribution $\Phi(\theta)$. For this purpose, we introduce a technique from evolutionary quantitative genetics, first analyzed by Lande (1976).[9] The approach offers a tractable method to study an evolutionary process within a continuously heterogeneous population. In particular, it will provide us with a parameter that is easy to interpret – the mean value of θ – that characterizes the distribution $\Phi(\theta)$ in an evolutionary equilibrium. In Section 5, we will discuss the applicability of this technique to our problem and the differences to standard replicator dynamics (Weibull, 1995).

Consider a large population which is heterogeneous along one trait α. The trait value is normally distributed with mean $\bar{\alpha}$ and variance σ^2. To simplify notation, we write $F(\alpha)$

[9] See Falconer and Mackay (1995) and Roff (1997) for an introduction to quantitative genetics.

for the cdf $F(\alpha, \bar{\alpha}, \sigma^2)$ and the density function is denoted by $f(\alpha)$. Let the fitness of an α-type, i.e., an individual with a trait value α, for a given distribution with mean $\bar{\alpha}$ be given by $w(\alpha, \bar{\alpha})$. Allowing individual fitness to depend on the distribution accounts for *frequency dependent* fitness. Fitness is called frequency dependent if the fitness of an α-individual also depends on the composition of the population.[10] In economic terms, frequency dependence is given if one group of agents – respectively the strategy played by these individuals – creates an externality on other agents' fitness.[11]

Within one generation, the change in the mean trait value in response to selection is defined as

$$\Delta\bar{\alpha} = \bar{\alpha}_s - \bar{\alpha}, \tag{9}$$

where $\bar{\alpha}_s$, the mean trait value after selection, is given by

$$\bar{\alpha}_s = \frac{1}{\bar{w}} \int \alpha w(\alpha, \bar{\alpha})\, dF(\alpha) \tag{10}$$

and $\bar{w}$, the mean fitness of the population, is

$$\bar{w} = \int w(\alpha, \bar{\alpha})\, dF(\alpha). \tag{11}$$

The selection described in (10) follows a replicator dynamic. While the initial frequency of a type was $f(\alpha)$, the post-selection frequency of this type, $\frac{w(\alpha,\bar{\alpha})}{\bar{w}} f(\alpha)$, will be higher for types with above-average fitness. Hence, in the computation of $\bar{\alpha}_s$, more successful types will get more weight than less successful types.

The analysis so far describes selection within one generation. In order to address the inter-generational evolution of the trait α, Lande (1976) introduces the following structure of reproduction. First, only selected individuals produce the next generation of offspring. Second, partner selection and genetic recombination transforms the post-selection distribution into an offspring distribution which is again normal: it is characterized by the

[10] As we will consider the variance to be fixed, we have suppressed this variable in $w(.)$ to ease notation.

[11] Consider, for instance, the decision to commit a crime where the likelihood of a criminal act to be 'successful' depends on the crime rate in the society. (E.g., the detection probability might be lower, the more other agents become criminals.) If decisions depend on individual risk preferences, the distribution of these preferences clearly influences the success of a criminal.

initial variance σ^2 but a different mean.[12] According to this structure, selection will then first lead to a distribution which deviates from the initial one. Starting from a norm distribution with mean $\bar{\alpha}$, the mean of the (non-normal) distribution after selection is given by $\bar{\alpha}_s$ from (10). After mating and reproduction, however, the distribution of α in the new generation is again normal with $F(\alpha, \bar{\alpha}_s, \sigma^2)$. While the variance is preserved, the mean of the distribution changes from $\bar{\alpha}$ to $\bar{\alpha}_s$. The direction of evolution is therefore determined by selection, characterized in (9) and (10). This allows us to analyze the evolutionary process in more detail.

From (11) we can derive the change in mean fitness from a marginal change in $\bar{\alpha}$,

$$\frac{\partial \bar{w}}{\partial \bar{\alpha}} = \int w(\alpha, \bar{\alpha}) \, \frac{\partial f(\alpha)}{\partial \bar{\alpha}} \, d\alpha + \int \frac{\partial w(\alpha, \bar{\alpha})}{\partial \bar{\alpha}} \, dF(\alpha). \tag{12}$$

While the first term characterizes the direct change in the mean fitness due to a change in the composition of the population, the second term depicts the indirect, frequency dependent fitness impact. From the density of the normal distribution we can easily compute $\partial f(\alpha)/\partial \bar{\alpha}$. Substituting in (12) and rearranging yields

$$\Delta \bar{\alpha} = \frac{1}{\bar{w}} \int w(\alpha, \bar{\alpha}) \, (\alpha - \bar{\alpha}) \; dF(\alpha) \tag{13}$$

(see Appendix A). The right-hand side in equation (13) characterizes pace and direction of the evolutionary process. As $\bar{w} > 0$ (per assumption), the direction of the evolutionary change in the mean trait value $\bar{\alpha}$ is determined by the sign of the integral in (13). Note that the integral term represents only the direct change in mean fitness (the first term in equation (12)). From (13) it therefore follows that the evolution of $\bar{\alpha}$ is independent of the frequency dependent fitness change associated with a change in $\bar{\alpha}$. If the direct fitness impact is positive (negative), the distribution will evolve towards a higher (lower) mean $\bar{\alpha}$. An *evolutionary equilibrium* is reached if $\Delta \bar{\alpha} = 0$. Such an equilibrium is characterized by

$$\int w(\alpha, \bar{\alpha}^e) \, (\alpha - \bar{\alpha}^e) \; dF(\alpha) = 0, \tag{14}$$

where $\bar{\alpha}^e$ denotes the mean trait value in equilibrium.

[12] The assumptions underlying this structure are justified by the observation that most metric traits have a normal distribution, or that the distribution can be transformed to normal by a change in the scale of measurement (e.g., by log transformation). Similar arguments are incurred to account for the independence of the variance in respect to the mean, and for that the variance is assumed constant over evolutionary time. Intuitively, mating among a large (selected) population would assure a constant variance as long as the mating process is random with respect to α (for a closer discussion, see Lande, 1976, Falconer and Mackay, 1995, and Roff, 1997). Admittedly, this case will be violated whenever assortative mating is based on the trait α.

4 Indirect Evolution of Conditional Cooperation

The method introduced in the previous Section is now applied to study the evolution of the distribution $\Phi(\theta)$ and the associated coevolution of cooperation in the model from Section 2. We interpret evolution as a cultural process, in the form of vertical norm transmission (e.g., education and socialization within families and peers).[13] Fitness measures the evolutionary success of a certain θ-type. The higher the relative success associated with a certain level of norm sensitivity, the more likely it is transmitted to the next generation. In this way, the evolutionary process endogenously shapes preferences. Individual behavior and thereby the level of cooperation within society evolves indirectly with the change in preferences from one generation to the next. The term 'generation' thereby describes a population with a given distribution of preferences.

We are convinced that the evolutionary success associated with a certain behavior is also determined by norm-based sanctions and rewards. Therefore we deviate from the typical approach in evolutionary economics, which considers economic payoffs as the sole determinants of evolutionary success (Fershtman and Weiss, 1998, Mengel, 2008). Apart from the payoff $y(x^i)$, success is also shaped by the norm-based sanctions $z(x^i, n)$. One might think of $z(.)$ as the objective costs for a norm violation stemming from social disapproval. These costs are non-increasing in the share of norm-violators n: in terms of evolutionary success, it is less costly to free-ride in a population where norm violations are widespread. For $n = 1$, the norm has completely eroded and norm violations have no consequences. The evolutionary success for an action x^i is then given by

$$w(x^i) = y(x^i) + z(x^i, n). \tag{15}$$

While $z(x^i, n)$ captures the objective costs of norm-based sanctions, the parameter θ measures the subjective sensitivity to theses sanctions – associated with heterogeneous levels of norm internalization. In terms of evolutionary success, $\theta = 1$ thus corresponds to the optimal internalization level (see below).[14]

The basic structure of the evolutionary process is the following. An initial generation with a given distribution $\Phi(\theta)$ faces the public good game described in Section 2. After the game is played, agents learn about the evolutionary success associated with different actions. According to the relative success, different θ-levels are then transmitted to the

[13] As highlighted by a referee, one might also argue that any heterogeneity in feelings of guilt, remorse, or other internal sanctions could be due to genetically determined personality traits that evolve in the very long rung.

[14] Note that $w(.)$ includes the payoff $y(.)$, which corresponds to the subjective (utility) costs from the public good provision. The utility formulation from (3) thus means that the two payoffs, that enter (15) additively, have different effects on utility (for $\theta \neq 1$).

next generation. The resulting change in the $\Phi(\theta)$ is assumed to be characterized by the process from (13). (In Section 5, we discuss the crucial differences of this approach from quantitative genetics to an adaptation process according to replicator dynamics.) This structure is studied for two scenarios. First, we consider the case, where each generation coordinates on one social equilibrium state n^*. Then we turn to the case, where – in the context of multiple equilibria – one generation will face different equilibrium states. We will call the first scenario a *homogenous* and the latter a *heterogeneous environment*.

4.1 Homogenous Environment

Let θ be normally distributed according to $\theta \sim \phi(\bar{\theta}, \sigma^2)$, and the cumulative distribution is given by $\Phi(\theta, \bar{\theta}, \sigma^2)$. Substituting for $y(x^i)$, $z(x^i, n)$ and $x^i = x(\theta^i, n)$ from (1), (2) and (5), we can express evolutionary success (15) as

$$w(\theta, \bar{\theta}) = \begin{cases} -c & \text{for } \theta > \hat{\theta}(n^*) \\ -s(n^*) & \text{for } \theta \leq \hat{\theta}(n^*) \end{cases} \tag{16}$$

where $n^* = \Phi(\hat{\theta}(n^*), \bar{\theta}, \sigma^2)$ is an a-stable equilibrium state as characterized by (7) and (8), for a normal distribution with mean $\bar{\theta}$ and an exogenous variance σ^2.

It is important to note three points here. First, it is only the heterogeneity in actions – determined by different levels of θ – which results in differences in evolutionary success. Within the group of cooperators or free-riders, the heterogeneity in θ does not result in different levels of evolutionary success. The evolutionary success thus relies on behavior which is in principle observable, rather than an unobservable θ. Second, evolutionary success as described by (16) is frequency dependent. As the distribution of θ changes, the share of free-riders n^* and thereby the costs of a norm deviation will change. Recall that the method introduced in Section 3 accounts for such spillovers. Third, we assume that a generation always coordinates on one equilibrium state n^*. In this sense, we study the evolution of norm sensitivities within a homogenous environment. Each new generation (with a new distribution of θ) is assumed to coordinate on an equilibrium state in the close neighborhood of the previous one – even if there exist multiple equilibrium states.[15]

[15]This assumption on equilibrium selection can be justified by the fact that after a small change in the distribution (i.e., in $\bar{\theta}$) there always exists a new, a-stable equilibrium state in the close neighborhood of the previous one if condition (8) holds with strict inequality. This close by equilibrium may be more salient than more distant equilibrium states and thus serves as a focal point.

The mean evolutionary success is defined by $\bar{w} = \int w(\theta, \bar{\theta})\, \phi(\theta)$.[16] Using (16), we obtain

$$\bar{w} = -c + (c - s(n^*)) \int_{-\infty}^{\hat{\theta}(n^*)} d\Phi(\theta) \tag{17}$$

with the integral expression being equal to $n^* = \Phi(\hat{\theta}(n^*), \bar{\theta}, \sigma^2)$. Following (13) and the steps described in section 3, one can easily characterize the intergenerational change in $\bar{\theta}$ (see Appendix A). For $0 < n^* < 1$, the 'direction' of the evolution is determined by

$$\text{sign}\left\{\Delta\bar{\theta}\right\} = \text{sign}\left\{s(n^*) - c\right\}. \tag{18}$$

This leads us to the following result:

Proposition 1 ***(i)*** *An evolutionary equilibrium where cooperators and free-riders coexist is characterized by $s(n^e) = c$, where $0 < n^e = \Phi(\hat{\theta}(n^e), \bar{\theta}^e, \sigma^2) < 1$ constitutes an a-stable equilibrium state, supported by a normal distribution with mean $\bar{\theta}^e$.* ***(ii)*** *In such an equilibrium, $\hat{\theta}(n^e) = 1$ and all agents have the same success $w(\theta, \bar{\theta}^e)$.* ***(iii)*** *An evolutionary equilibrium where cooperation fails, $n^{e1} = 1$, is characterized by an a-stable equilibrium state $n^{e1} = \Phi(\hat{\theta}(n^{e1}), \bar{\theta}^{e1}, \sigma^2)$, supported by a normal distribution with mean $\bar{\theta}^{e1}$.*

Proof. See Appendix B.

The evolutionary equilibrium n^e described in part (i) of the Proposition is characterized by a positive share of cooperators such that there is no differential between free-riders and cooperators with respect to evolutionary success. In equilibrium, the preferences of agents with $\theta^i = \hat{\theta}(n^e)$, who are indifferent between defection and cooperation, coincide with the evolutionary success as given by (15) since $\hat{\theta}(n^e) = 1$. In other words, these θ-types are 'perfectly adapted' – their individual norm sensitivity perfectly accounts for the evolutionary costs of a norm violation. In addition, there is also an evolutionary equilibrium where everybody free-rides. While we know from Lemma 1 that $n^* = 1$ constitutes a possible equilibrium state for *any* distribution, condition (8) has to hold to guarantee the asymptotic stability of the equilibrium state. Therefore, any level $\bar{\theta}$ for which (8) holds at $n^* = 1$ could be the mean of the distribution in an evolutionary equilibrium with zero cooperation, n^{e1}. By the time the whole society free-rides, the evolutionary pressure on $\bar{\theta}$ to decline vanishes and the system reaches a rest point.[17]

[16] In the following, we will assume $\bar{w} > 0$ which can be assured, e.g., by adding a sufficiently large constant payoff to $w(.)$, without changing any of our results.

[17] In principle, we could also describe an evolutionary equilibrium with full cooperation. However, an equilibrium state with $n^* = 0$ would only be supported by a distribution with $\bar{\theta} \to \infty$. We do not include this case in our analysis, as such a distribution would violate $\theta \in \mathbb{R}$.

Let us now turn to the existence of these different types of equilibria.

Proposition 2 ***(i)*** *Iff $s(0) > c$, there exists an evolutionary equilibrium with $0 < n^e < 1$.* ***(ii)*** *For all distributions fulfilling (8) at $n^* = 1$, there exists an evolutionary equilibrium with $n^{e1} = 1$.* ***(iii)*** *If $c > s(0)$ and (8) holds for $n^* = 1$, this is the only equilibrium.*

Proof. See Appendix B.

The result from Proposition 2 is straightforward. If the costs of cooperating are higher than the loss from a norm violation even for the state where $n^* = 0$, free-riding yields a higher evolutionary success than cooperation for any n. Starting from any $n^* < 1$, the evolutionary process induces $\bar{\theta}$ to fall and society moves towards an equilibrium with full defection, $n^{e1} = 1$. However, if sanctions are sufficiently strong such that cooperators have a higher evolutionary success than free-riders for the full-cooperation state $n^* = 0$, there exists an equilibrium state $0 < n^e < 1$ where both actions yield the same evolutionary success.

Finally, we address the evolutionary stability of the system. Note that we apply two stability concepts. In Section 2 we focused on stability *within* one generation (a-stability) which – for a given distribution of θ – requires an equilibrium state to be robust to small behavioral trembles. Evolutionary stability now demands that preferences remain stable *between* generations. If this is the case, small mistakes in the transmission of norms will not affect the equilibrium. We call an evolutionary equilibrium locally *evolutionary stable (e-stable)* if $d\Delta\bar{\theta}/d\bar{\theta} < 0$ holds in the close neighborhood of $\bar{\theta}^e$ (or $\bar{\theta}^{e1}$). Consider, for example, a positive shock on $\bar{\theta}$. One can derive from (7) that an increase in the mean norm sensitivity would result in a drop in the share of free-riders below n^e. The stability condition would then demand that $\Delta\bar{\theta} < 0$ which provides a pressure on $\bar{\theta}$ to fall and consequently on n^* to increase, thereby adapting 'back' towards the initial equilibrium $\bar{\theta}^e$ or n^e. In turns out that an evolutionary equilibrium where cooperators and free-riders coexist is never e-stable.

Proposition 3 *An evolutionary equilibrium with $0 < n^e < 1$ is never e-stable. In contrast, an evolutionary equilibrium with $n^{e1} = 1$ is locally e-stable.*

Proof. See Appendix B.

Due to assumption A1, $s'(n) \leq 0$. Hence, any small deviation from n^e would tip the balance in evolutionary success between the two strategies. After a positive shock on $\bar{\theta}^e$, the share of free-riders falls short of n^e and we get $s(n) \geq c$. Cooperators would be more successful than free-riders, $\bar{\theta}$ would increase and n^* would decline further. If, on the

other hand, the level of free-riding exceeds n^e, the norm-based sanctions would become less effective and we get $c \geq s(n)$. Free-riders, i.e., individuals with low values of θ, are evolutionarily more successful than cooperators. Consequently $\bar{\theta}$ decreases and the system moves into an equilibrium with $n^{e1} = 1$. Note that the system would return to such an equilibrium n^{e1} after small shocks in $\bar{\theta}$, as in the neighborhood of $n^{e1} = 1$ there holds $c > s(n^{e1})$ due to A1. Hence, an evolutionary equilibrium with $\bar{\theta}^{e1}$ and n^{e1} would be stable.

The analysis provided so far yields an unsatisfactory result. While there can exist an evolutionary equilibrium where free-riders and cooperators coexist, such an equilibrium turns out to be unstable. The system either evolves towards an equilibrium where the norm has eroded and everybody free-rides, or the society would evolve towards full cooperation. Despite the fact that losses follow the condition pattern $s(n)$, the evolution of norm sensitivities within a homogenous environment does not support conditional cooperators.

4.2 Heterogeneous Environment

So far, we have considered a homogenous environment. Agents encounter one particular situation – one equilibrium state – and evolution shapes their preferences according to the strength of the social norm in this equilibrium. In reality, however, we often face heterogeneous environments: people are guided by norms against littering or against crime, when they walk through clean and littered parks, through nice and run down neighborhoods (Funk, 2005, Glaeser et al., 1996); smokers might have a no-smoking norm in mind when they are at parties at which people smoke, but also at those where nobody smokes (Nyborg and Rege, 2003); we might work in a firm where many co-workers cheat but also face projects where others' exert high efforts (Ichino and Maggi, 2000). In the following, we discuss a stylized framework which capture such heterogeneous environments.[18] In contrast to the case of a homogenous environment, we find (potentially) e-stable evolutionary equilibria where cooperators and free-riders coexist.

Consider an initial distribution such that assumption A2(ii) is fulfilled. In this case, there exists a multiplicity of equilibria (see Lemma 1). Within each generation, the population sometimes coordinates on an a-stable equilibrium state n_a^*, sometimes on n_b^* with $n_j^* = \Phi(\hat{\theta}(n_j^*), \bar{\theta}, \sigma^2)$ for $j \in \{a, b\}$. Without loss of generality, we assume $n_a^* < n_b^*$. The likelihood with which a generation coordinates on equilibrium state n_j^* is exogenously given

[18] One might argue that different outcomes are simply due to population heterogeneity, e.g., different distributions of θ in 'good' and 'bad' environments. While this heterogeneity obviously exists, it only partially explains the diversity of observed behavior (see, e.g., the discussion in Glaeser et al., 1996). Our analysis therefore abstracts from heterogeneity in population characteristics.

by $0 < \pi_j < 1$.[19] The actions an agent i with θ^i chooses according to (5) in the equilibrium states n_a^*, respectively n_b^*, is denoted by (x_a^i, x_b^i). The evolutionary success for (x_a^i, x_b^i) is now given by

$$w\left(x_a^i, x_b^i\right) = \sum_{j=a,b} \pi_j \left(y(x_j^i) + z(x_j^i, n_j^*)\right). \tag{19}$$

From $n_a^* < n_b^*$ and (6) follows $\hat{\theta}(n_a^*) < \hat{\theta}(n_b^*)$. Hence, we will observe three different strategies: on the one hand, agents with $\theta^i \leq \hat{\theta}(n_a^*)$ will free-ride in both equilibrium states. Agents with $\theta^i > \hat{\theta}(n_b^*)$ on the other hand, will cooperate in both states. A third group of individuals, those with $\hat{\theta}(n_a^*) < \theta^i \leq \hat{\theta}(n_b^*)$, behaves conditionally cooperative. They cooperate in equilibrium state a, where many others cooperate as well, but defect in state b, as more of the others free-ride. Making use of (1), (2) and (5), we can express the evolutionary success of agents with different norm sensitivities in the following way:

$$w(\theta, \bar{\theta}) = \begin{cases} -c & \text{for } \theta > \hat{\theta}(n_b^*) \\ -\pi_a c - \pi_b s(n_b^*) & \text{for } \hat{\theta}(n_a^*) < \theta \leq \hat{\theta}(n_b^*) \\ -\pi_a s(n_a^*) - \pi_b s(n_b^*) & \text{for } \theta \leq \hat{\theta}(n_a^*) \end{cases} \tag{20}$$

The crucial difference to the case of a homogenous environment is the fact that agents with intermediate levels of θ obtain fitness payoffs from two different actions. The success of the conditionally cooperative strategy consists of the cooperation payoff for equilibrium state a plus the payoff from free-riding in state b.

Using (20) we can compute the mean evolutionary success of the population for a given π_a and $\pi_b = 1 - \pi_a$,

$$\bar{w} = -c + \pi_a \left(c - s(n_a^*)\right) \int_{-\infty}^{\hat{\theta}(n_a^*)} d\Phi(\theta) + (1 - \pi_a) \left(c - s(n_b^*)\right) \int_{-\infty}^{\hat{\theta}(n_b^*)} d\Phi(\theta). \tag{21}$$

From (13) one can easily show that the evolution of $\bar{\theta}$ is now determined by $\Delta\bar{\theta} = \frac{1}{\bar{w}}\Psi$ with

$$\Psi \equiv \pi_a \left(s(n_a^*) - c\right) \left(\bar{\theta} n_a^* - \bar{\theta}_a^*\right) + (1 - \pi_a) \left(s(n_b^*) - c\right) \left(\bar{\theta} n_b^* - \bar{\theta}_b^*\right), \tag{22}$$

[19] At this point, one could extend the analysis in several directions. One could derive the likelihood π_j endogenously from the relative size of the basin of attraction for a particular equilibrium state n_j^*. Moreover, one could easily consider heterogenous environments with more than two stable equilibrium states. These extensions do not effect our main results.

and $\bar{\theta}_j^*$ captures the mean level of θ among the free-riders for equilibrium state n_j^*.[20] The evolutionary dynamics on $\bar{\theta}$ are now given by

$$\text{sign}\left\{\Delta\bar{\theta}\right\} = \text{sign}\left\{\Psi\right\} \tag{23}$$

This leads to the following proposition:

Proposition 4 ***(i)*** *An evolutionary equilibrium in a heterogeneous environment is characterized by $\Psi = 0$, where the stable equilibrium states $n_a^e = \Phi(\hat{\theta}(n_a^e), \bar{\theta}^e, \sigma^2)$ and $n_b^e = \Phi(\hat{\theta}(n_b^e), \bar{\theta}^e, \sigma^2)$ are supported by a normal distribution with mean $\bar{\theta}^e$.* ***(ii)*** *If $n_b^e < 1$, there holds $s(n_a^e) > c > s(n_b^e)$.*

Proof. See Appendix B.

The Proposition characterizes an evolutionary equilibrium for a heterogeneous environment. As long as $n_b^e < 1$, the distribution in the evolutionary equilibrium supports two equilibrium states such that $s(n_a^e) > c > s(n_b^e)$.[21] In terms of evolutionary success, cooperation dominates free-riding in equilibrium state a. For state b, however, the opposite holds: free-riding is more widespread, and the losses from violating the norm are lower than the costs of cooperation. This implies

Corollary 1 *In an evolutionary equilibrium in a heterogeneous environment with $n_b^e < 1$ conditional cooperators are evolutionarily more successful than free-riders and cooperators.*

Proof. From Proposition 4(ii) we know that $s(n_a^e) > c > s(n_b^e)$. Using this in (20) proves the Corollary. ■

Figure 2 graphically illustrates an example of such an evolutionary equilibrium. The graph on the left-hand side captures a system with a distribution $\Phi(\theta)$ and a function $\hat{\theta}(n)$ supporting two stable equilibrium states $n_a^* < n_b^* < 1$. The graph on the right-hand side depicts the difference in evolutionary payoffs between the strategies for the two equilibria.

Figure 2 about here

[20] The definition is analogous to (A.10) in the Appendix. Equivalently, the derivation of $\Delta\bar{\theta}$ and Ψ is analogous to the one of (A.9), discussed in Appendix A.

[21] Another possible equilibrium would be $n_b^e = 1$ and $s(n_a^e) = c$. As this type of equilibrium has similar properties to the one discussed in the previous section, we do not discuss this case. Moreover, the equilibrium condition, $\Psi = 0$, would also be fulfilled for (i) $n_a^e = 0$ and $n_b^e = 1$ as well as for (ii) $n_a^e = 0$ and $n_b^e < 1$ with $s(n_b^e) = c$. Note, however, that assumption A1 implies $\hat{\theta}(0) > 0$. Unless $\bar{\theta} \to \infty$, there is always a positive mass of individuals with $\theta \leq \hat{\theta}(0)$ which makes an equilibrium state $n_a^e = 0$ impossible.

From Figure 2 and from the analysis above (compare Proposition 2) it is clear that $s(0) > c$ is a necessary condition for an evolutionary equilibrium to exist. In addition, assumption A2(ii) has to hold in order to guarantee a multiplicity of equilibria. Analogous to before, the necessary and sufficient condition for the local e-stability of an evolutionary equilibrium is $d\Delta\bar{\theta}/d\bar{\theta} < 0$. From this we derive

Proposition 5 *Sufficient conditions for the e-stability of an evolutionary equilibrium with* $n_b^e < 1$ *are given by* $n_a^e \leq \min\{\gamma_a; \delta_a\}$ *and* $\gamma_b \leq n_b^e \leq \delta_b$, *with*

$$\gamma_j \equiv \int_{-\infty}^{\hat{\theta}(n_j^e)} \phi(\theta) \frac{\left(\theta - \bar{\theta}^e\right)^2}{\sigma^2} d\theta,$$

$$\delta_j \equiv \frac{\bar{\theta}_j^*}{\bar{\theta}^e} + \phi(\hat{\theta}(n_j^e))\ \hat{\theta}(n_j^e)\left(1 - \hat{\theta}(n_j^e)\right)\left(1 - \frac{\hat{\theta}(n_j^e)}{\bar{\theta}^e}\right).$$

Proof. See Appendix B.

As it is difficult to discuss the intuition behind the stability conditions, we conducted a series of numerical simulations.[22] Typically, we found two levels of $\bar{\theta}$ which supported an evolutionary equilibrium. The one with the higher mean norm-sensitivity was *always* stable, even for cases where the (sufficient) condition $n_a^e \leq \min\{\gamma_a; \delta_a\}$ from Proposition 5 was violated. We are therefore confident that stable evolutionary equilibria within a heterogeneous environment exist for a wide range of parameters. Our conjecture is backed by a straightforward intuition: small shocks in the norm transmission would not change the result from Corollary 1 – conditional cooperation would still perform more successfully than the two unconditional strategies. Since conditional cooperators have intermediate values of θ, preferences in the 'middle' of the θ distribution are evolutionarily more successful and dominate against those with more extreme (either low or high) θ-values.

The evolutionary dominance of conditional cooperators is the main result of our analysis. Individuals who lack pro-social preferences (those with low θ values) as well as individuals with 'overly' pro-social preferences (very high values of θ) play one particular strategy, irrespectively of the other agents' behavior. In a stable evolutionary equilibrium within a homogenous environment, one of these two strategies will dominate the other. In a heterogeneous environment, i.e., when individuals face a 'good' state with high levels of cooperation and a 'bad' state with widespread free-riding, a third strategy appears:

[22] In the Appendix, we analyze the stability conditions in more detail and show that they can both be fulfilled. In the simulations, we mainly work with the functional form $s(n) = \lambda(1 - r(n^a/a - n^b/b))$ and parameters in the range $c = 1$, $\lambda \in (1, 2]$, $r \in [1.5, 2.5]$, $a \in [1, 2]$, $b \in [2, 4]$, a standard deviation $\sigma \in [1.5, 2.5]$ and $\pi_a \in (0, 1)$. The program code is available from the authors upon request.

conditional cooperation. In the adaptation to such a heterogeneous environment, the two unconditional strategies prove less successful that the conditional strategy. Agents who cooperate in the good but free-ride in the bad state dominate the free-riders in the former and the cooperators in the latter environment. Thus, the evolutionary pressure to adapt to heterogeneous environments provides a simple explanation for the success of conditionally cooperative behavior.[23]

5 Discussion

5.1 Replicator Dynamics

Would our results still hold if evolution followed a conventional replicator dynamic rather than the quantitative genetic process? Consider a population with $N \to \infty$ possible values of θ, indexed by $\ell \in \{1, ..., N\}$, where an N-dimensional vector $g = [g^\ell]$ depicts the distribution of θ (compare Bisin et al., 2009). Let the frequency of a type, g^ℓ, evolve according to

$$\dot{g}^\ell = g^\ell \left(w(\theta^\ell) - \bar{w} \right). \tag{24}$$

From the analysis in Section 4.1 it immediately follows that any distribution which supports an equilibrium share n^e with $s(n^e) = c$ also constitutes an evolutionary equilibrium according to (24). If $s(n^e) = c$ holds, neither free-riders nor cooperators have any evolutionary advantage (compare Proposition 1) and we would get $w(\theta^\ell) = \bar{w} \Rightarrow \dot{g}^\ell = 0$ for all ℓ. Similarly, the (a-)stability properties of such an equilibrium with $0 < n^e < 1$ carry over: any small deviation from n^e would either lead to a breakdown in cooperation or a move towards full cooperation.

The analysis of Section 4.2 suggests that conditional cooperation will always dominate the two unconditional strategies in a heterogeneous environment. This holds for *any* evolutionary dynamics. Adaptation according to (24), however, would eliminate preferences that induce unconditional strategies. In an evolutionary equilibrium, the whole population would consist of conditional cooperators. Everybody would cooperate in one equilibrium state ($n_a^* = 0$), and free-ride in the other ($n_b^* = 1$). Any distribution of θ with $g^\ell \geq 0$ for $\hat{\theta}(0) \leq \theta^\ell \leq \hat{\theta}(1)$ and $g^\ell = 0$ otherwise which supports these equilibrium states would constitute an evolutionary equilibrium. Hence, the dynamics from (24) do not (in general) lead to a society with one homogenous level of norm sensitivity.

[23] The intuition also applies for alternative behavioral models that capture conditional cooperation (Fehr and Schmidt, 2006). Models of self-centered or group-centered inequality aversion, for instance, predict unconditional free-riding (unconditional cooperation) for low (high) but conditional cooperation for intermediate values of inequality aversion.

5.2 Quantitative Genetics

The evolutionary model developed by Lande (1976) considers a normally distributed trait that evolves according to the relative fitness differential $w(\theta)/\bar{w}$: the frequency of θ-type with an evolutionary payoff above (below) the mean will increase (shrink). The resulting non-normal distribution is transformed back to a normal distribution with a new mean. In this vein, the mean trait value $\bar{\theta}$ evolves whereas the normal character and the variance of the distribution are preserved. Our motivation to apply this method is mainly technical. On the one hand, the method provides a tractable tool to study the evolution of a continuous type distribution within the basic model from Section 2. On the other hand, the approach seems perfectly suited to study a process with frequency dependence, i.e., externalities in the different types' evolutionary success created by the change in the type distribution.

Admittedly, the quantitative genetic method has several limitations.[24] Within our framework, it implies an imperfect evolutionary process, as the initial variance in θ is maintained during the course of evolution. Hence, by using this method we neglect the case where all agents adapt one unique θ value (e.g., $\theta = 1$). One could justify this implication by systematic disturbances that are embedded in any norm transmission process.[25] In the context of a cultural process, this appears more plausible than the case of perfect adaptation to one unique level of norm sensitivity. Even if, e.g., the parents want to 'install' a certain level of θ among their offspring, other peer influences might affect the actually transmitted level.

Note that a perfectly homogenous population would (in general) not constitute a stable evolutionary equilibrium according to the dynamic from (24) either. In contrast to Lande's approach, however, the process from (24) does not allow for a co-existence of different strategies, i.e., free-riding, cooperation and conditional cooperation, in an evolutionary equilibrium within a heterogeneous environment. The behavioral heterogeneity that emerges in the equilibrium characterized in Proposition 4 is simply a consequence of the constant variance. For the case of a normal distribution with an infinitesimally small variance, however, the evolutionary equilibrium would consist of a population of conditional cooperators (such that $n_a^* \to 0$ and $n_b^* \to 1$). For this special case, behavior (but not necessarily the distribution of θ) in the evolutionary equilibrium would be equivalent for replicator dynamics and the quantitative genetic approach.

[24] One crucial limitation of the method would be the case with evolutionary pressure on low *and* high θ-types to grow. This would suggest an evolution towards a bimodal distribution which is excluded by assumption in Lande's approach. Such a disruptive evolution cannot occur in our set-up.

[25] If the errors in the norm transmission are normally distributed – i.e., if the influences in the education and socialization of the next generation contain random elements – and remain constant during evolution, these deviations from perfect adaptation in θ would maintain a normal distribution $\Phi(\theta)$. This would also apply for a genetically determined heterogeneity in θ (see footnote 13).

6 Conclusion

While the impact of heterogeneous 'habitats' on evolutionary processes is well studied by biologists,[26] this idea has so far been neglected in evolutionary economics. In this paper, we take a first step towards closing this gap in the literature. We develop a model of voluntary public good provisions in the context of a social norm for cooperation. As the strength of social norms depends on the level of cooperation, there is scope for multiple equilibria. Society may coordinate on an equilibrium with a high level of cooperation, where norm deviations would result in significant sanctions, or on a state with widespread free-riding and weak norm-enforcement. We link this multiplicity of equilibria to the idea of heterogeneous habitats, in the sense that the evolutionary success of a certain norm-sensitivity, and the behavior induced by it, is evaluated for different equilibria of the game. Following an indirect evolutionary approach, individual norm sensitivities are shaped endogenously according to their performance in equilibrium states with a strong norm and states with a weak norm. In such heterogeneous environments, conditional cooperation is more successful than any unconditional strategy. In the 'cooperative' environment, conditional cooperators follow the norm and avoid the loss that free-riders incur from violating a strong norm. In the environment where the norm is weak and norm-based sanctions hardly play a role, conditional cooperators reap the same payoff as free-riders, which dominates that of an unconditional cooperator. Hence, the preferences underlying conditional cooperation are well adapted to heterogeneous environments. An intermediate level of norm sensitivity allows individuals to react flexibly to different social situations. Thereby, they dominate unconditional strategies which are specialized on one particular condition.

Members of modern human societies typically interact in various cooperation problems where cooperation fails sometimes, but works quite well in other situations. We face both cooperative and uncooperative environments, clean public parks and littered ones, projects where co-workers exert high efforts and those in which others shirk. Our analysis suggests that exactly such a heterogeneity in our social environments is a driving force in the evolutionary success of conditional cooperation.

Acknowledgements

We would like to thank two anonymous referees, Florian Herold, Simon Gächter, Aljaž Ule and Seminar Participants at CREED, University of Amsterdam, for helpful comments. Traxler acknowledges financial support by the European Network for the Advancement of Behavioural Economics (ENABLE). The usual disclaimer applies.

[26] Among many others, see Levins (1968), Maynard Smith and Hoekstra (1980).

Appendix

A.1 – Section 3

For the density of the normal distribution, $f(\alpha)$, one can easily derive

$$\frac{\partial f(\alpha)}{\partial \bar{\alpha}} = f(\alpha)\frac{\alpha - \bar{\alpha}}{\sigma^2}. \tag{A.1}$$

Making use of this term in (12) and rearranging, we get

$$\frac{\partial \bar{w}}{\partial \bar{\alpha}} = \frac{1}{\sigma^2}\int \left[\alpha\, w(\alpha,\bar{\alpha})\, f(\alpha) - \bar{\alpha}\, w(\alpha,\bar{\alpha})\, f(\alpha)\right] d\alpha + \int \frac{\partial w(\alpha,\bar{\alpha})}{\partial \bar{\alpha}} dF(\alpha). \tag{A.2}$$

From (10), respectively (11), it follows that the first expression in the first integral equals $\bar{\alpha}_s \bar{w}$, and the second expression is $\bar{\alpha}\bar{w}$. We arrive at

$$\frac{\partial \bar{w}}{\partial \bar{\alpha}} = \frac{\bar{w}}{\sigma^2}\left(\bar{\alpha}_s - \bar{\alpha}\right) + \int \frac{\partial w(\alpha,\bar{\alpha})}{\partial \bar{\alpha}} dF(\alpha). \tag{A.3}$$

Rearranging and making use of (9) yields

$$\Delta\bar{\alpha} = \frac{\sigma^2}{\bar{w}}\left(\frac{\partial \bar{w}}{\partial \bar{\alpha}} - \int \frac{\partial w(\alpha,\bar{\alpha})}{\partial \bar{\alpha}} dF(\alpha)\right) \tag{A.4}$$

which is equivalent to

$$\Delta\bar{\alpha} = \frac{\sigma^2}{\bar{w}}\int w(\alpha,\bar{\alpha})\,\frac{\partial f(\alpha)}{\partial \bar{\alpha}}\, d\alpha. \tag{A.5}$$

Substituting for (A.1) we arrive at (13).

A.2 – Section 4

The mean evolutionary success is given by

$$\bar{w} = -s(n^*)\int_{-\infty}^{\hat{\theta}(n^*)} d\Phi(\theta) - c\int_{\hat{\theta}(n^*)}^{\infty} d\Phi(\theta). \tag{A.6}$$

As $\Phi(\hat{\theta}(n^*)) = n^*$, we can rearrange $\bar{w}$ and get

$$\bar{w} = -\left(1 - n^*\right)c - n^* s(n^*). \tag{A.7}$$

From this follows (17).

As we have demonstrated in section 3, only the direct impact of a change in $\bar{\theta}$ is important for the evolution of this variable. The indirect effect that arises since the equilibrium state n^* changes with any change in the distribution (which, in turn, affects $s(n)$, the 'frequency dependent' element determining evolutionary success) is irrelevant. Hence, we can follow (13) and obtain

$$\Delta\bar{\theta} = \frac{1}{\bar{w}} \{ - \int_{-\infty}^{\hat{\theta}(n^*)} s(n^*) \left(\theta - \bar{\theta}\right) d\Phi(\theta) - \int_{\hat{\theta}(n^*)}^{\infty} c \left(\theta - \bar{\theta}\right) d\Phi(\theta) \}. \tag{A.8}$$

Adding $[\int_{-\infty}^{\hat{\theta}} c(\theta - \bar{\theta}) d\Phi(\theta) - \int_{-\infty}^{\hat{\theta}} c(\theta - \bar{\theta}) d\Phi(\theta)] = 0$, rearranging terms and noting that $c \int_{-\infty}^{\infty} (\theta - \bar{\theta}) d\Phi(\theta) = 0$, we get

$$\Delta\bar{\theta} = \frac{1}{\bar{w}} \left(s(n^*) - c\right) \left(\bar{\theta} n^* - \bar{\theta}^*\right) \tag{A.9}$$

where $\bar{\theta}^*$ represents the mean level of θ among the n^* agents who free-ride in equilibrium,

$$\bar{\theta}^* \equiv \int_{-\infty}^{\hat{\theta}(n^*)} \theta d\Phi(\theta). \tag{A.10}$$

As long as $0 < n^* < 1$, there holds $\bar{\theta} n^* > \bar{\theta}^*$. As $\bar{w} > 0$ per assumption, we arrive at (18).

B – Proofs

Proof of Lemma 1. As we can rewrite condition (7) as $\Phi^{-1}(n^*) - \hat{\theta}(n^*) = 0$, it follows immediately from A1 and A2(i) that there always exists an equilibrium with $n^* = 1$. From A1 we know $s(0) > 0 \Rightarrow \hat{\theta}(0) > 0$ which implies $\hat{\theta}(n) > \Phi^{-1}(n)$ for $n \to 0$. From this follows that A2(ii) assures that there must exist at least one $n^* \in (0, 1)$ where $\Phi^{-1}(n^*) = \hat{\theta}(n^*)$ holds, since both $\hat{\theta}(n)$ and $\Phi^{-1}(n)$ are continuously increasing functions defined over the unit interval. ■

Proof of Proposition 1. The proof of **(i)** follows immediately from (A.9). From (4) we know that $c = \hat{\theta}(n^*) s(n^*)$ must hold for any equilibrium state. $s(n^e) = c$ then implies $\hat{\theta}(n^e) = 1$. Using this in (16) and substituting for (4) proves **(ii)**. Part **(iii)** derives from $n^* = 1 \Rightarrow \bar{\theta} n^* = \bar{\theta}^*$. Hence, for $n^{e1} = 1$ the term in the last brackets in (A.9) is zero and $\Delta\bar{\theta} = 0$. ■

Proof of Proposition 2. **(i)** Since $c > s(n)$ for $n \to 1$ and $s(.)$ is continuously non-increasing in n, $s(0) > c$ assures that there exists a level of n where $s(n) = c$ holds. Moreover, we can always find a distribution $\phi(\theta, \bar{\theta}, \sigma^2)$, a function $s(n)$ and a level c, which supports such an equilibrium share of free-riders n^e. **(ii)** From Lemma 1 we know that $n^* = 1$ is supported by any distribution as long as A1 and A2(i) hold. Proposition 1(iii) implies that any equilibrium with $n^* = 1$ where (8) holds, constitutes an evolutionary equilibrium n^{e1}. **(iii)** From A1 follows that $c > s(0)$ implies $c > s(n)$ for all $n \in [0, 1]$. It therefore follows from $c > s(0)$ that there cannot exist an equilibrium with $n^e < 1$, as $\nexists\, n$ with $s(n) = c$. ■

Proof of Proposition 3. From (A.9) one can derive

$$\begin{aligned}
\frac{d\Delta\bar{\theta}}{d\bar{\theta}} &= \frac{1}{\bar{w}}\left(s(n^*) - c\right) \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta) - \phi(\theta)\frac{\left(\theta - \bar{\theta}\right)^2}{\sigma^2} d\theta \\
&\quad -\frac{1}{\bar{w}^2}\left[\frac{\partial \bar{w}}{\partial \bar{\theta}} + \frac{\partial \bar{w}}{\partial n^*}\frac{\partial n^*}{\partial \bar{\theta}}\right]\left(s(n^*) - c\right) \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta)\left(\bar{\theta} - \theta\right) d\theta \qquad \text{(A.11)} \\
&\quad +\frac{1}{\bar{w}}\left[s'(n^*) \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta)\left(\bar{\theta} - \theta\right) d\theta + \left(s(n^*) - c\right)\frac{\partial \hat{\theta}(n^*)}{\partial n^*}\phi(\hat{\theta})\left(\bar{\theta} - \hat{\theta}(n^*)\right)\right]\frac{\partial n^*}{\partial \bar{\theta}}
\end{aligned}$$

where we made use of the Leibniz Rule of integral differentiation to derive the last term in the third line's squared brackets. Rearranging and making use of (4), (7) and (A.10), we arrive at

$$\begin{aligned}
\frac{d\Delta\bar{\theta}}{d\bar{\theta}} &= \frac{1}{\bar{w}}\left(s(n^*) - c\right)\left[n^* - \int_{-\infty}^{\hat{\theta}(n^*)} \phi(\theta)\frac{\left(\theta - \bar{\theta}\right)^2}{\sigma^2} d\theta\right] \\
&\quad -\frac{1}{\bar{w}^2}\left[\frac{\partial \bar{w}}{\partial \bar{\theta}} + \frac{\partial \bar{w}}{\partial n^*}\frac{\partial n^*}{\partial \bar{\theta}}\right]\left(s(n^*) - c\right)\left(\bar{\theta}n^* - \bar{\theta}^*\right) \qquad \text{(A.12)} \\
&\quad +\frac{1}{\bar{w}}\left[\left(\bar{\theta}n^* - \bar{\theta}^*\right) + \left(s(n^*) - c\right)\frac{\hat{\theta}(n^*)}{s(n^*)}\phi(\hat{\theta})\left(\hat{\theta}(n^*) - \bar{\theta}\right)\right] s'(n^*)\frac{\partial n^*}{\partial \bar{\theta}}
\end{aligned}$$

From Proposition 1 we know that an evolutionary equilibrium with $0 < n^e < 1$ is characterized by $s(n^e) = c$. Therefore, the expressions in the first and the second line of (A.12) equal zero for such an equilibrium n^e. Using (7), one can easily show that $\partial n^*/\partial \bar{\theta} \leq 0$ for any stable equilibrium state n^*. As $s'(n^*) \leq 0$ and $\bar{\theta}n^* > \bar{\theta}^*$ for $0 < n^* < 1$, it follows that the expression in the third line of (A.12) must be non-negative and we get $d\Delta\bar{\theta}/d\bar{\theta} \geq 0$ for

any evolutionary equilibrium with $0 < n^e < 1$. Such an evolutionary equilibrium is never stable.

Let us now consider an evolutionary equilibrium with $n^{e1} = 1$. Since $\hat{\theta}(n^{e1}) \to \infty$ for $n^{e1} = 1$, the integral term in the first line of (A.12) equals the variance σ^2 and the term in the squared brackets becomes zero. For $n^{e1} = 1$ there also holds $\bar{\theta} n^* = \bar{\theta}^*$ and the expression in the second line of (A.12) also equals zero. From $s(n^{e1}) = 0$, $\hat{\theta}(n^{e1}) \to \infty$ and $\bar{\theta} n^* = \bar{\theta}^*$, it follows that the term in the third line's squared brackets is strictly negative. Together with $\partial n^*/\partial \bar{\theta} \leq 0$ and $s'(n^*) \leq 0$ this implies that $d\Delta\bar{\theta}/d\bar{\theta} < 0$ holds for $n^{e1} = 1$. ■

Proof of Proposition 4. Part **(i)** follows immediately from (23). Part **(ii)** derives from (22): Note that $\bar{\theta} n_j^* > \bar{\theta}_j^*$ as long as $n_j^* < 1$. Hence, the first term in (22) would be negative if $c > s(n_a^e)$. Since $n_a^e < n_b^e$, (6) implies that the second term would be negative as well. We would get $\Psi < 0$. Therefore $c > s(n_a^e)$ cannot hold in an equilibrium with $n_b^e < 1$. Iff $s(n_a^e) > c$, the first term in (22) is positive. In order to get $\Psi = 0$ for $n_b^e < 1$, the second term in (22) must be negative, which holds for $c > s(n_b^e)$. ■

Proof of Proposition 5. Analogously to (A.12) we can derive from (21) and (22)

$$\begin{aligned}
\frac{d\Delta\bar{\theta}}{d\bar{\theta}} \;=\; & \frac{1}{\bar{w}} \sum_j \pi_j \left(s(n_j^*) - c \right) \left[n_j^* - \int_{-\infty}^{\hat{\theta}(n_j^*)} \phi(\theta) \frac{\left(\theta - \bar{\theta}\right)^2}{\sigma^2} d\theta \right] \\
& - \frac{1}{\bar{w}^2} \left[\frac{\partial \bar{w}}{\partial \bar{\theta}} + \sum_j \pi_j \frac{\partial \bar{w}}{\partial n_j^*} \frac{\partial n_j^*}{\partial \bar{\theta}} \right] \Psi \qquad \text{(A.13)} \\
& + \frac{1}{\bar{w}} \sum_j \pi_j \left[\bar{\theta} n_j^* - \bar{\theta}_j^* + \left(s(n_j^*) - c \right) \frac{\hat{\theta}(n_j^*)}{s(n_j^*)} \phi(\hat{\theta}(n_j^*)) \left(\hat{\theta}(n_j^*) - \bar{\theta} \right) \right] s'(n_j^*) \frac{\partial n_j^*}{\partial \bar{\theta}}
\end{aligned}$$

Since in an evolutionary equilibrium $\Psi = 0$ (Proposition 4), the second line of (A.13) equals zero. In an equilibrium as characterized in Proposition 4(ii), i.e., where $n_b^e < 1$, it holds that $s(n_a^e) > c > s(n_b^e)$. If the squared bracket term in the first line is positive for equilibrium state n_b^e and negative for n_a^e, the expression in the first line of (A.13) is unambiguously negative. The two corresponding conditions are

$$n_a^e \leq \int_{-\infty}^{\hat{\theta}(n_a^e)} \phi(\theta) \frac{\left(\theta - \bar{\theta}^e\right)^2}{\sigma^2} d\theta, \qquad \text{and} \qquad n_b^e \geq \int_{-\infty}^{\hat{\theta}(n_b^e)} \phi(\theta) \frac{\left(\theta - \bar{\theta}^e\right)^2}{\sigma^2} d\theta. \qquad \text{(A.14)}$$

(Note that the integral term in (A.14) takes values in the range $(0, 0.5]$ for $0 < n_a^e \leq 0.5$ and $[0.5, 1)$ for $0.5 \leq n_a^e < 1$.)

Let us now turn to the third line of (A.13). Remember that $s'(n_j^*) \leq 0$ and $\partial n_j^*/\partial \bar{\theta} \leq 0$ since both equilibrium states n_j^* are stable as characterized by (8). It is therefore sufficient for the expression in the third line to be negative, if the term in the squared brackets is negative for both equilibrium states. Rearranging, we get the following condition

$$n_j^e \leq \frac{\bar{\theta}_j^*}{\bar{\theta}} + \phi(\hat{\theta}(n_j^e))\ \hat{\theta}(n_j^e) \left(1 - \hat{\theta}(n_j^e)\right) \left(1 - \frac{\hat{\theta}(n_j^e)}{\bar{\theta}}\right), \tag{A.15}$$

where we have substituted for (4). The first term on the RHS of (A.15) is positive for any $n^* > 0$. Moreover, for $n_a^* \leq 0.5$ there holds $\hat{\theta}(n_a^*) \leq \bar{\theta}$. Since $1 - \hat{\theta}(n_j^*) = (s(n_j^*) - c)/\ s(n_j^*)$, $s(n_a^*) > c$ implies that the second term on the RHS is also positive for $n_a^* \leq 0.5$. For an equilibrium state $n_b^* \geq 0.5$ we know that $\hat{\theta}(n_b^*) \geq \bar{\theta}$. From $s(n_b^*) < c$ then follows that the RHS is again strictly positive. (As the first term approaches unity for $n_b^* \to 1$ and since the second term is strictly positive, the RHS of (A.15) could be strictly larger than unity for high levels of n_b^*. For $n_a^* \to 0$, the second term will be positive, as $\hat{\theta}(0) > 0$ holds due to assumption A1. Hence, condition (A.15) should hold for extreme levels of n_j^*.) ■

References

Akerlof, George A. (1980), A Theory of Social Custom, of which Unemployment may be one Consequence, *The Quarterly Journal of Economics* 94(4), 749-775.

Azar, Ofer H. (2005), The Social Norm of Tipping: Does it Improve Social Welfare? *Journal of Economics* 85(2), 141-173.

Benabou, Roland and Jean Tirole (2006), Incentives and Prosocial Behavior, *American Economic Review* 96(5), 1652-1678.

Bisin, Alberto, Giorgio Topa and Thierry Verdier (2009), Cultural transmission, socialization and the population dynamics of multiple-trait distributions, *International Journal of Economic Theory* 5(1), 139-154.

Coleman, James S. (1990), *Foundations of Social Theory*, Harvard University Press, Cambridge (MA).

Corneo, Giacomo (1997), The theory of the open shop trade union reconsidered, *Labour Economics* 4(1), 71-84.

Corneo, Giacomo and Olivier Jeanne (1997), Snobs, bandwagons, and the origin of social customs in consumer behavior, *Journal of Economic Behavior and Organization* 32(3), 333-347.

Elster, Jon (1998), Emotions and Economic Theory, *Journal of Economic Literature* 36(1), 47-74.

Falconer, Douglas S. and Trudy F.C. Mackay (1995), *Introduction to Quantitative Genetics*. 4th Edition. Addison Wesley Longman, New York.

Fehr, Ernst and Klaus Schmidt (2006), The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories, in: Serge-Christophe Kolm and Jean Mercier Ythier (Eds.), *Handbook on the Economics of Giving, Reciprocity and Altruism*, Vol.1, North Holland, Amsterdam.

Fershtman, Chaim and Yoram Weiss (1998), Social Rewards, Externalities and Stable Preferences, *Journal of Public Economics* 70(1), 53-73.

Fischbacher, Urs, Simon Gächter and Ernst Fehr (2001), Are People Conditionally Cooperative? Evidence from a Public Goods Experiment, *Economics Letters* 71(3), 397-404.

Frey, Bruno and Stephan Meier (2004), Social Comparisons and Pro-social Behavior: Testing 'Conditional Cooperation' in a Field Experiment, *American Economic Review* 94(5), 1717-1722.

Funk, Patricia (2005), Governmental Action, Social Norms, and Criminal Behavior, *Journal of Institutional and Theoretical Economics* 127(3), 522-535.

Gächter, Simon (2007), Conditional Cooperation: Behavioral Regularities from the Lab and the Field and their Policy Implications, in: Bruno S. Frey and Alois Stutzer (Eds.), *Economics and Psychology. A Promising New Cross-Disciplinary Field*, CESifo Seminar Series, MIT Press.

Güth, Werner (1995), An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives, *International Journal of Game Theory* 24(4), 323-344.

Güth, Werner and Menahem E. Yaari (1992), Explaining reciprocal behavior in simple strategic games: An evolutionary approach, in: Ulrich Witt (Ed.), *Explaining process and change: Approaches to evolutionary economics*, Michigan University Press, Ann Arbor.

Glaeser, Edward L., Bruce Sacerdote and Jose A. Scheinkman (1996), Crime and Social Interactions, *The Quarterly Journal of Economics*, 111(2), 507-548.

Ichino, Andrea and Giovanni Maggi (2000), Work Environment And Individual Background: Explaining Regional Shirking Differentials In A Large Italian Firm, *The Quarterly Journal of Economics* 115(3), 1057-1090.

Keser, Claudia and Frans van Winden (2000), Conditional Cooperation and Voluntary Contributions to Public Goods, *Scandinavian Journal of Economics* 102, 23-39.

Lande, Russel (1976), Natural selection and random genetic drift in phenotypic evolution, *Evolution* 30(2), 314-334.

Levins, Richard (1968), *Evolution in changing environments*. Princeton University Press, Princeton.

Lindbeck, Assar, Sten Nyberg and Jörgen W. Weibull (1999), Social Norms and Economic Incentives in the Welfare State, *The Quarterly Journal of Economics* 114(1), 1-35.

Maynard Smith, John and Rolf Hoekstra (1980), Polymorphism in a varied environment: how robust are the models? *Genetical Research* 35, 45-57.

Mengel, Friederike (2008), Matching structure and the cultural transmission of social norms, *Journal of Economic Behavior and Organization* 67(3-4), 608-623.

Nyborg, Karine and Mari Rege (2003), On Social Norms: The Evolution of Considerate Smoking Behavior, *Journal of Economic Behavior and Organization* 52(3), 323-340.

Roff, Derek A. (1997), *Evolutionary Quantitative Genetics*. Chapman and Hall, New York.

Traxler, Christian and Joachim Winter (2009), Survey Evidence on Conditional Norm Enforcement, Max Planck Institute for Research on Collective Goods, Working Paper 2009/03.

Weibull, Jorgen W. (1995), *Evolutionary Game Theory*. The MIT Press, Cambridge (MA).

Figure 1

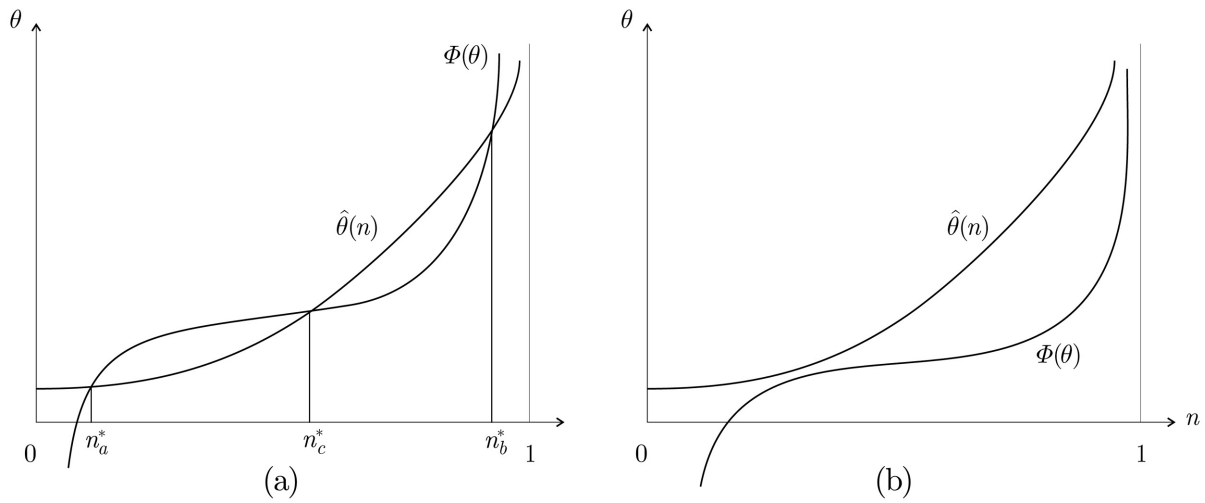
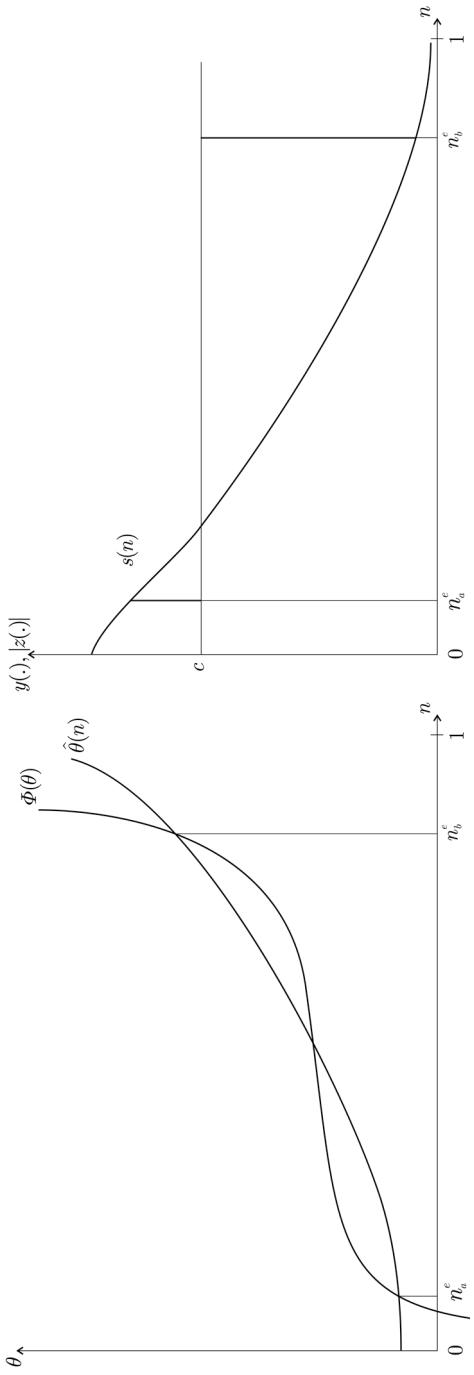

θ
$\Phi(\theta)$
$\hat{\theta}(n)$
n_a^*
n_c^*
n_b^*
0
1
(a)
θ
$\hat{\theta}(n)$
$\Phi(\theta)$
n
0
1
(b)

Figure 2


n
1
n_b^e
$s(n)$
$y(\cdot), |z(\cdot)|$
c
0
n_a^e
n
1
$\hat{\theta}(n)$
$\Phi(\theta)$
n_b^e
θ
0
n_a^e