

# Why the rich are nastier than the poor –

## A note on optimal punishment

Steffen Huck and Wieland Müller  
*Humboldt–University\**

### Abstract

Studying evolutionarily successful behavior we show in a general framework that when individuals maximizing payoff differentials invest resources in punishing others. Interestingly, these investments are increasing in individuals' own wealth and decreasing in the wealth of others.

### 1 Introduction

Experimental evidence suggests that many individuals have a preference for punishing others even though these punishments inflict costs on themselves (see e.g. Bolton and Zwick (1995) or Abbink *et al.* (1996) who provide very strong support for this punishment hypothesis). Huck and Oechssler (1995) show that a preference for 'revenge' will typically be stable in the context of ultimatum games which implies that resources are almost always split equally. Sethi and Somanathan (1996) show that punishments can be evolutionarily successful in the context of common resource games since they help to establish cooperative behavior.

In this paper we show in a more general framework that costly punishment of others can be evolutionarily profitable. The intuition of this result is simple. If a punitive action harms others more than oneself, one can increase one's relative payoff by carrying out this action, and since evolution is not driven by absolute but by relative payoffs such actions may be evolutionarily successful. Moreover, we show that under plausible assumptions optimal punishment is increasing in the personal wealth of the one who punishes and decreasing in the wealth of the one who is punished.

While the term 'punishment' is often understood as describing a reciprocal action—someone who was treated in an unkind way responds by harming his opponent—we do not restrict the analysis to this kind of negative reciprocity. More generally, we consider any actions harming others no matter why the actions are carried out. Thus, the analysis covers not only negative reciprocity but also acts of nastiness or malevolence.

The results from this analysis are quite robust. In fact, we consider two different frameworks—a rather simple one in Section 2 in which only one individual can carry out a punitive action (lowering the *average* payoff of all others) and a more complicated (but also more plausible) one in Section 3 in which all individuals have the opportunity to punish all others (lowering the *individual* payoffs of others). In both sections we establish qualitatively similar results.

These results can serve as explanations for some real life phenomena which is discussed in the concluding Section 4.

---

\*Institute for Economic Theory III, Spandauer Strasse 1, 10178 Berlin, Germany, Fax +49 30 20935706, email [huck@wiwi.hu-berlin.de](mailto:huck@wiwi.hu-berlin.de).

## 2 Model A: One against all

Consider a situation in which all individuals of a population have gained some consumable resources endowing them with a certain absolute material payoff. The allocation may be the result of a move of nature, the result of a game played by all individuals, the result of many games played by subgroups of the population, the result of a market process, or the result of anything else. Suppose that given such a situation an individual can carry out an action harming others by investing some of his resources. Of course, rational actors whose preferences only rely on absolute material payoffs would never carry out such actions. However, studies of preference evolution provide clear-cut evidence that evolution does not yield types being rational in that ‘Friedman style.’<sup>1</sup>

It was Alchian (1950) who first pointed out that this cannot be expected since evolution is driven by *payoff differentials*. In this study we are interested in evolutionarily successful *behavior*. Therefore, we analyse which behavioral consequences are to be expected when individuals maximize payoff differentials rather than absolute payoffs.<sup>2</sup> This analysis will yield *optimal punishment profiles* which can be seen as the behavioral analogon to evolutionarily successful preferences.

First some notation.

Let  $M_i$  be individual  $i$ ’s material payoff in the first phase, let  $\overline{M}_{-i}$  be the average material payoff of all others and assume that both variables are observable. A punishment profile is described by a function  $p_i : \mathbb{R}^2 \rightarrow \mathbb{R}$ , prescribing for every vector  $(M_i, \overline{M}_{-i})$  a punitive action  $p_i = p_i(M_i, \overline{M}_{-i})$ . When individual  $i$  has carried out a punishment his final payoff is denoted by  $f(M_i, p_i)$ , and the average final payoff of the others by  $g(\overline{M}_{-i}, p_i)$ . We make the following straightforward assumptions about  $f$  and  $g$ :

### Assumptions:

- a) Both functions are twice continuously differentiable. Furthermore,  $\frac{\partial f}{\partial M_i} > 0$  and  $\frac{\partial g}{\partial \overline{M}_{-i}} > 0$ , i.e. the final payoff is increasing in material wealth gained in the first phase.
- b)  $\frac{\partial f}{\partial p_i} < 0$  and  $\frac{\partial g}{\partial p_i} < 0$ , i.e. the final payoff is decreasing in  $i$ ’s punishment.
- c)  $\frac{\partial^2 f}{\partial p_i^2} < 0$  and  $\frac{\partial^2 g}{\partial p_i^2} > 0$ , i.e. while the marginal cost of punishments is increasing for individual  $i$ , the absolute marginal effect of punishment on others is decreasing.
- d)  $\frac{\partial^2 f}{\partial p_i \partial M_i} > 0$ , i.e. the marginal cost of punishment is decreasing in personal wealth.

While assumptions a) and b) purely reflect the definition of punishments, assumptions c) and d) impose restrictions which, however, seem very natural.

To make model A as simple as possible we assume furthermore that at each point in time (i.e. after each ‘allocation phase’) only one individual selected by chance has the opportunity to punish. In model B this assumption will be replaced.

Now suppose that individuals maximize the payoff differential<sup>3</sup>  $f(M_i, p_i) - g(\overline{M}_{-i}, p_i)$ . As we will show this implies that one can derive an optimal punishment profile  $p^*$ .

To compute  $p^*$  one has to solve the following problem:

---

<sup>1</sup>Friedman (1953) argued that evolutionary forces would bring about such preferences. But his conjecture turned out to be false (see e.g. De Long, Shleifer, Summers, and Waldmann 1990 or Blume and Easley 1992 and 1995).

<sup>2</sup>For a related approach see Akerlof (1976).

<sup>3</sup>There is a broad class of evolutionary dynamics (including the well-known replicator dynamics) in which growth rates of types depend monotonically on this expression. This illustrates well that types who maximize this expression have the best chances to survive and spread in an evolutionary process.

$$\begin{aligned} \text{Maximize } R_i(M_i, \overline{M}_{-i}, p_i) &= f(M_i, p_i) - g(\overline{M}_{-i}, p_i) \text{ w.r.t. } p_i \\ \text{subject to } M_i &\geq p_i \geq 0 \end{aligned} \quad (1)$$

This yields the following lemma:

**Lemma 1** *The optimal punishment profile is characterized by the implicit function  $p^*(M_i, \overline{M}_{-i})$  :  $\frac{\partial g(\overline{M}_{-i}, p_i)}{\partial p_i} - \frac{\partial f(M_i, p_i)}{\partial p_i} = 0$  if  $\partial g(\overline{M}_{-i}, 0)/\partial p_i < \partial f(M_i, 0)/\partial p_i$  and by  $p_i^* = 0$  otherwise.*

**Proof** The first order condition for maximization of  $R_i$  is

$$\frac{\partial R_i}{\partial p_i} = 0 \Leftrightarrow \frac{\partial g(\overline{M}_{-i}, p_i)}{\partial p_i} - \frac{\partial f(M_i, p_i)}{\partial p_i} = 0. \quad (2)$$

Due to assumption c)  $R_i$  is concave in  $p_i$ . Note that there exists some  $\hat{p}_i$  such that  $\frac{\partial R_i}{\partial p_i} < 0$  for all  $p_i > \hat{p}_i$ . Therefore, equation (2) has a (unique) solution with  $p_i > 0$  if and only if

$$\partial g(\overline{M}_{-i}, 0)/\partial p_i < \partial f(M_i, 0)/\partial p_i. \quad (3)$$

If (3) does not hold, this implies that  $\frac{\partial R_i}{\partial p_i} < 0$  for all  $p_i \geq 0$ .  $\square$

With the help of Lemma 1 we can prove the following proposition showing that the optimal punishment profile  $p^*$  implies that the amount of punishment (or nastiness) is increasing in the material payoff gained in the pre-punishment phase.

**Proposition 2**  $\partial g(\overline{M}_{-i}, 0)/\partial p_i < \partial f(M_i, 0)/\partial p_i \Leftrightarrow \frac{\partial p_i^*(M_i, \overline{M}_{-i})}{\partial M_i} > 0$ .

**Proof** Note first that if (3) holds for some  $M_i = M'$  it also holds for all  $M_i > M'$ . If it holds let  $H(M_i, \overline{M}_{-i}, p_i)$  be the implicit function defined by 2. Applying the implicit function theorem yields  $\frac{\partial p_i^*(M_i, \overline{M}_{-i})}{\partial M_i} = -\frac{\partial H/\partial M_i}{\partial H/\partial p_i} = \frac{-\partial^2 f/\partial p_i \partial M_i}{\partial^2 f/\partial p_i^2 - \partial^2 g/\partial p_i^2} > 0$ . If (3) does not hold,  $\frac{\partial p_i^*(M_i, \overline{M}_{-i})}{\partial M_i} = 0$ .  $\square$

While this result may be taken as an evolutionary justification for ‘why the rich are nastier than the poor’ we next show that under an additional assumption which has some plausibility our approach may also explain ‘why the poor are treated nastier than the rich’.

**Proposition 3** *If an individual punishes at all and if  $\frac{\partial^2 g}{\partial p_i \partial \overline{M}_{-i}} > 0$ , then  $\frac{\partial p_i^*(M_i, \overline{M}_{-i})}{\partial \overline{M}_{-i}} < 0$ .*

**Proof** Note first that if (3) holds for some  $\overline{M}_{-i} = M''$  it also holds due to assumption d) for all  $\overline{M}_{-i} < M''$ . Now let  $H$  be defined as before. Then,  $\frac{\partial p_i^*(M_i, \overline{M}_{-i})}{\partial \overline{M}_{-i}} = -\frac{\partial H/\partial \overline{M}_{-i}}{\partial H/\partial p_i} = \frac{-\partial^2 g/\partial p_i \partial \overline{M}_{-i}}{\partial^2 f/\partial p_i^2 - \partial^2 g/\partial p_i^2} < 0$ .  $\square$

There is a German idiom which tries to summarize the ‘stylized fact’ that in many conflicts those who are worst off in the beginning are those who loose most in conflicts since they are treated most badly. Our result shows that such behavioral patterns are evolutionarily profitable if the absolute marginal effects of punishments are decreasing in the victim’s wealth. In this case it is more effective to harm somebody who is down already than somebody who is on the top.

### 3 Model B: All against all

In the following we will consider a model similar to the above one but with the exception that after each allocation phase all individuals can punish all individuals, i.e. each individual  $i$  is allowed to carry out a variety of punitive actions directed to other *specific* individuals. This makes some additional notation necessary.

Let  $M_i$  be the private wealth of individual  $i$ . Let  $p_{ih}$  ( $h \neq i$ ) be the amount of wealth<sup>4</sup> individual  $i$  invests in harming individual  $h$ . Let furthermore  $S_i = \sum_{h \neq i} p_{ih}$  and  $T_i = \sum_{h \neq i} p_{hi}$ , i.e.  $S_i$  denotes the total amount of wealth individual  $i$  invests in punishments, and  $T_i$  denotes the total amount of wealth which is invested by others to punish  $i$ . Let  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$  be the function capturing the final material payoff with  $f$  depending on  $M_i$ ,  $S_i$ , and  $T_i$ . Finally, let  $\mathcal{N} := \{1, 2, \dots, n\}$  be the set of individuals. A punishment profile is now characterized by a mapping  $p : \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$  which prescribes for given material wealth of all individuals a vector of punitive actions  $p_i = (p_{i1}, \dots, p_{ii-1}, p_{ii+1}, \dots, p_{in})$ .

Assuming similar preferences as above we can denote the objective function of individual  $i$  as

$$R_i(M, S, T) = f(M_i, S_i, T_i) - \frac{1}{n} \sum_{j=1}^n f(M_j, S_j, T_j)$$

where  $M = (M_1, \dots, M_n)$ ,  $S = (S_1, \dots, S_n)$ ,  $T = (T_1, \dots, T_n)$ .

Concerning the function  $f$  we make the following assumptions:

**Assumptions:**

- a) The function  $f$  is twice continuously differentiable in all entries. Furthermore,  $\frac{\partial f}{\partial M_i} > 0$ .
- b)  $\frac{\partial f}{\partial S_i} < 0$  and  $\frac{\partial f}{\partial T_i} < 0$ , i.e. final payoff is decreasing in punishments.
- c)  $\frac{\partial^2 f}{\partial S_i^2} < 0$  and  $\frac{\partial^2 f}{\partial T_i^2} > 0$ , i.e. while the marginal cost of punishments is increasing, the absolute marginal effect of punishment on others is decreasing.
- d)  $\frac{\partial^2 f}{\partial S_i \partial M_i} > 0$ , i.e. the marginal cost of punishment is decreasing in personal wealth.

To derive the optimal punishment profile we must now consider the effects of *interaction* at the second stage. The question is which punitive actions are optimal when also all others can carry out punishments. This means nothing but solving the second stage as a game in which all individuals maximize  $R_i$ .

First of all we show that this subgame (the second phase) *has* an equilibrium. To do this we have to solve the following problem simultaneously for all  $i \in \mathcal{N}$ :

$$\begin{aligned} \text{Maximize} \quad & R_i(M, S, T) = f(M_i, S_i, T_i) - \frac{1}{n} \sum_{j=1}^n f(M_j, S_j, T_j) \quad \text{w.r.t. } p_i \\ \text{subject to} \quad & \sum_{h \neq i} p_{ih} \leq M_i \\ & \text{and } p_{ih} \geq 0 \text{ for all } h \in \mathcal{N}_{-i} \end{aligned} \tag{4}$$

where  $\mathcal{N}_{-i} := \mathcal{N} \setminus \{i\}$ .

Now let the strategy space of individual  $i$  be  $P_i := \left\{ p_i \in \mathbb{R}_+^{n-1} : \sum_{h \neq i} p_{ih} \leq M_i \right\}$  which is a nonempty, compact and convex subset of the Euclidean space  $\mathbb{R}^{n-1}$ . Because of the assumptions

---

<sup>4</sup>Here, we take already for granted that self-punishment is never optimal.

made above the payoff function  $R_i$  of each individual  $i$  is continuous in all entries and concave in  $p_i$ . Therefore, we can apply a well-known existence theorem (see e.g. Theorem 1.2 in Fudenberg and Tirole 1991, p. 34) and conclude that for each  $M = (M_1, \dots, M_n) \in \mathbb{R}_+^n$  system (4) has a solution  $p^* = (p_1^*, \dots, p_n^*) \in \mathbb{R}_+^{n(n-1)}$  which because of the concavity of the functions  $R_i$  is unique, i.e. due to the theorem  $p^*$  is the unique pure-strategy Nash equilibrium of the second phase subgame. Note that each  $p_i^*$  ( $i \in \mathcal{N}$ ) is itself a vector, i.e.  $p_i^* = (p_{i1}^*, \dots, p_{ii-1}^*, p_{ii+1}^*, \dots, p_{in}^*)$  consisting of the punitive actions executed by individual  $i$  in equilibrium.

In the following we will establish results analogous to those of model A.

Imagine a situation where (given  $M \in \mathbb{R}_+^n$ ) each individual except  $i$  (which is fixed in the following) chooses his actions according to the Nash solution, e.g. only individual  $i$  is left to choose a strategy in reaction to what the others do. Thus, only individual  $i \in \mathcal{N}$  has to solve the maximization problem (4). The according Kuhn–Tucker conditions can be written as follows:

$$\frac{\partial R_i}{\partial p_{ih}} + \lambda \leq 0 \quad p_{ih} \geq 0 \quad \text{and} \quad p_{ih} \left( \frac{\partial R_i}{\partial p_{ih}} + \lambda \right) = 0 \quad \text{for all } h \neq i \quad (5)$$

$$M_i - \sum_{h \neq i} p_{ih} \geq 0 \quad \lambda \geq 0 \quad \text{and} \quad \lambda \left( M_i - \sum_{h \neq i} p_{ih} \right) = 0. \quad (6)$$

This is a set of simultaneous conditions which determine the optimal strategy for individual  $i$  that is known to exist. Assume that we have an inner solution  $p_i^*$ , i.e.  $p_{ih}^* > 0$  for all  $h \in \mathcal{N}_{-i}$  and  $\sum_{h \neq i} p_{ih} < M_i$  which (because of (6)) implies that  $\lambda = 0$ . Then, according to (5) the following simultaneous equations must hold:

$$(n-1) \frac{\partial f(M_i, S_i, T_i)}{\partial S_i} - \frac{\partial f(M_h, S_h, T_h)}{\partial T_h} = 0 \quad \text{for all } h \in \mathcal{N}_{-i}.$$

Denoting the left hand sides of these equations with  $F_i^h$  for all  $h \neq i$  we can write down the system<sup>5</sup>

$$F_i^h(p_{i1}, \dots, p_{ii-1}, p_{ii+1}, \dots, p_{in}; M_1, \dots, M_n) = 0 \quad \text{for all } h \neq i. \quad (7)$$

For obvious reasons let us call  $p_{ih}$  ( $h \neq i$ ) the *endogenous* variables and  $M_i$  ( $i \in \mathcal{N}$ ) the *exogenous* variables. First of all from Theorem 1.2 in Fudenberg and Tirole (1991) we know that the above system has the solution  $p_i^* = (p_{i1}^*, \dots, p_{ii-1}^*, p_{ii+1}^*, \dots, p_{in}^*)$ , i.e. the point  $(p_i^*, M)$  satisfies (7). Second, due to our assumptions all  $F_i^h$  have continuous partial derivatives with respect to all variables. In order to apply the implicit function theorem we have to check that for the Jacobian,  $|J_i|$ , of the endogenous variables  $p_{i1}, \dots, p_{ii-1}, p_{ii+1}, \dots, p_{in}$  of system (7) it is true that  $|J_i| \neq 0$  at the point  $(p_i^*, M)$ .

Let us agree upon the following notation:  $f^j := f(M_j, S_j, T_j)$ ,  $j \in \mathcal{N}$ .

**Remark 1** According to assumptions a) and c) for the endogenous-variable Jacobian,  $|J_i|$ , it holds that

---

<sup>5</sup>Note that the functions  $F_i^h$  also depend on  $p_{jk}^*$ ,  $j \in \mathcal{N}_{-i}$ ,  $k \in \mathcal{N}_{-j}$ , i.e. the punitive actions in equilibrium of all other individuals  $j \neq i$  which we consider as parameters in the following analysis and which are therefore skipped.

$$\begin{aligned}
|J_i(p_i^*, M)| &= \begin{vmatrix} \frac{\partial F_i^1}{\partial p_{i1}} & \cdots & \frac{\partial F_i^1}{\partial p_{ii-1}} & \frac{\partial F_i^1}{\partial p_{ii+1}} & \cdots & \frac{\partial F_i^1}{\partial p_{in}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_i^{i-1}}{\partial p_{i1}} & \cdots & \frac{\partial F_i^{i-1}}{\partial p_{ii-1}} & \frac{\partial F_i^{i-1}}{\partial p_{ii+1}} & \cdots & \frac{\partial F_i^{i-1}}{\partial p_{in}} \\ \frac{\partial F_i^{i+1}}{\partial p_{i1}} & \cdots & \frac{\partial F_i^{i+1}}{\partial p_{ii-1}} & \frac{\partial F_i^{i+1}}{\partial p_{ii+1}} & \cdots & \frac{\partial F_i^{i+1}}{\partial p_{in}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_i^n}{\partial p_{i1}} & \cdots & \frac{\partial F_i^n}{\partial p_{ii-1}} & \frac{\partial F_i^n}{\partial p_{ii+1}} & \cdots & \frac{\partial F_i^n}{\partial p_{in}} \end{vmatrix}_{(p_i^*, M)} \\
&= \left. \begin{aligned} &(-1)^{n+1} \prod_{j \in \mathcal{N}_{-i}} \frac{\partial^2 f^j}{\partial T_j^2} \\ &+ (-1)^n (n-1) \frac{\partial^2 f^i}{\partial S_i^2} \sum_{r \in \mathcal{C}_{\mathcal{N}_{-i}}^{n-2}} \frac{\partial^2 f^{r_1}}{\partial T_{r_1}^2} \frac{\partial^2 f^{r_2}}{\partial T_{r_2}^2} \cdots \frac{\partial^2 f^{r_{n-1}}}{\partial T_{r_{n-1}}^2} \end{aligned} \right\} \begin{aligned} &< 0 \text{ for } n \text{ even} \\ &> 0 \text{ for } n \text{ odd,} \end{aligned}
\end{aligned}$$

where  $\mathcal{C}_{\mathcal{N}_{-i}}^{n-2}$  is the set of all combinations (without repetition) of order  $n-2$  of the set  $\mathcal{N}_{-i}$ .

**Proof:** See the appendix.

Note that this statement holds true not only at the point  $(p_{-i}^*, M)$  but due to our assumptions at all points  $(p_i, M)$ ,  $p_i \in P_i$ ,  $M \in \mathbb{R}_+^n$ .

Thus all conditions of the implicit function theorem are fulfilled and we can write

$$p_{ih}^* = G_{ih}(M_1, \dots, M_n) \quad \text{for all } h \in \mathcal{N}_{-i}$$

emphasizing that in equilibrium the punitive actions (the endogenous variables) of individual  $i$  are implicit functions of the initial values of wealth (the exogenous variables) of all individuals.

Our main intention is to look at comparative static implications of this general framework. More precisely, we are interested whether an increase of an exogenous variable such as  $M_i$  (or  $M_h$ ,  $h \neq i$ ) results in an increase or a decrease of the equilibrium value of the punitive action  $p_{ih}^*$ , i.e. whether individual  $i$  is going to punish individual  $h$  ( $h \neq i$ ) harder or not in equilibrium if the material wealth of individual  $i$  (or  $h$ ) increases. By applying again the implicit function theorem together with Cramer's Rule (see Chiang 1984, pp. 210) we can write down the relevant partial derivative for the first statement as:

$$\frac{\partial p_{ih}^*(M_1, \dots, M_n)}{\partial M_i} = \frac{|J_{ih}(p_{-i}^*, M)|}{|J_i(p_{-i}^*, M)|}$$

where  $|J_{ih}|$  is simply the endogenous-variable Jacobian  $|J_i|$  with the  $h$ -th column replaced by the vector

$$-\left( \frac{\partial F_i^1}{\partial M_i}, \dots, \frac{\partial F_i^{i-1}}{\partial M_i}, \frac{\partial F_i^{i+1}}{\partial M_i}, \dots, \frac{\partial F_i^n}{\partial M_i} \right)^T = -(n-1) \frac{\partial^2 f^i}{\partial S_i \partial M_i} (1, \dots, 1, 1, \dots, 1)^T \quad (8)$$

which is evaluated at the equilibrium  $(p_i^*, M)$ .

**Remark 2** According to assumptions a), c) and d) it holds that

$$|J_{ih}(p_{-i}^*, M)| = \left. \begin{aligned} &(-1)^{n-1} (n-1) \frac{\partial^2 f^i}{\partial S_i \partial M_i} \prod_{\substack{j \in \mathcal{N}_{-i} \\ j \neq h}} \frac{\partial^2 f^j}{\partial T_j^2} \end{aligned} \right\} \begin{aligned} &< 0 \text{ for } n \text{ even} \\ &> 0 \text{ for } n \text{ odd.} \end{aligned}$$

**Proof:** See the appendix.

We are now ready to state the first result:

**Proposition 4** For each  $i \in \mathcal{N}$  and each  $h \in \mathcal{N}_{-i}$  we have  $\frac{\partial p_{ih}^*(M_1, \dots, M_n)}{\partial M_i} > 0$ .

**Proof:** According to the Implicit Function Theorem, *Remark 1* and *2* and using Cramer's Rule we have

$$\begin{aligned} \frac{\partial p_{ih}^*(M_1, \dots, M_n)}{\partial M_i} &= \frac{|J_{ih}(p_i^*, M)|}{|J_i(p_i^*, M)|} \\ &= \frac{(-1)^{n-1}(n-1) \frac{\partial^2 f^i}{\partial S_i \partial M_i} \prod_{\substack{j \in \mathcal{N}_{-i} \\ j \neq h}} \frac{\partial^2 f^j}{\partial T_j^2}}{(-1)^{n+1} \prod_{j \in \mathcal{N}_{-i}} \frac{\partial^2 f^j}{\partial T_j^2} + (-1)^n (n-1) \frac{\partial^2 f^i}{\partial S_i^2} \sum_{r \in \mathcal{C}_{\mathcal{N}_{-i}}^{n-2}} \frac{\partial^2 f^{r_1}}{\partial T_{r_1}^2} \frac{\partial^2 f^{r_2}}{\partial T_{r_2}^2} \dots \frac{\partial^2 f^{r_{n-1}}}{\partial T_{r_{n-1}}^2}} > 0. \quad \square \end{aligned}$$

This result is analogous to Proposition 2 derived in the previous section. Again it shows that those who are rich will invest more in punishments than those who are poor. Next we turn to the question how the equilibrium value of  $p_{ih}^*$  varies if  $M_h$  changes. The relevant partial derivative is now

$$\frac{\partial p_{ih}^*(M_1, \dots, M_n)}{\partial M_h} = \frac{|J'_{ih}(p_i^*, M)|}{|J_i(p_i^*, M)|}$$

where  $|J'_{ih}|$  is again the endogenous-variable Jacobian  $|J_i|$  with the  $h$ -th column replaced by the vector

$$-\left( \frac{\partial F_i^1}{\partial M_h}, \dots, \frac{\partial F_i^{h-1}}{\partial M_h}, \frac{\partial F_i^h}{\partial M_h}, \frac{\partial F_i^{h+1}}{\partial M_h}, \dots, \frac{\partial F_i^n}{\partial M_h} \right)^T = \frac{\partial^2 f^h}{\partial T_h \partial M_h} (0, \dots, 0, 1, 0, \dots, 0)^T \quad (9)$$

which again is evaluated at the equilibrium  $(p_i^*, M)$ .

**Remark 3** According to assumptions *a)* and *c)* it holds that for all  $h \in \mathcal{N}$

$$\begin{aligned} \text{sgn} \left( |J'_{ih}(p_i^*, M)| \right) &= \\ &= \text{sgn} \left( (-1)^{n+1} \frac{\partial^2 f^h}{\partial T_h \partial M_h} \left( - \prod_{j \in \mathcal{N}_{-i, h}} \frac{\partial^2 f^j}{\partial T_j^2} + (n-1) \frac{\partial^2 f^i}{\partial S_i^2} \right) \times \sum_{r \in \mathcal{C}_{\mathcal{N}_{-i, h}}^{n-3}} \frac{\partial^2 f^{r_1}}{\partial T_{r_1}^2} \frac{\partial^2 f^{r_2}}{\partial T_{r_2}^2} \dots \frac{\partial^2 f^{r_{n-3}}}{\partial T_{r_{n-3}}^2} \right) \\ &= (-1)^n \cdot \text{sgn} \left( \frac{\partial^2 f^h}{\partial T_h \partial M_h} \right) \end{aligned}$$

**Proof:** See the appendix.

**Proposition 5** For each  $i \in \mathcal{N}$  and each  $h \in \mathcal{N}_{-i}$  we have  $\text{sgn} \left( \frac{\partial p_{ih}^*(M_1, \dots, M_n)}{\partial M_h} \right) = -\text{sgn} \left( \frac{\partial^2 f^h}{\partial T_h \partial M_h} \right)$ .

**Proof:** According to the Implicit Function Theorem, Remark 1 and 3 and using Cramer's Rule

$$\begin{aligned}
& \text{we have } \operatorname{sgn} \left( \frac{\partial p_{ih}^*(M_1, \dots, M_n)}{\partial M_h} \right) = \operatorname{sgn} \left( \frac{J'_{ih}(p_i^*, M)}{|J_i(p_i^*, M)|} \right) = \\
& = \operatorname{sgn} \left( \frac{(-1)^{n+1} \frac{\partial^2 f^h}{\partial T_h \partial M_h} \left( - \prod_{j \in \mathcal{N}_{-i, h}} \frac{\partial^2 f^j}{\partial T_j^2} + (n-1) \frac{\partial^2 f^i}{\partial S_i^2} \sum_{r \in \mathcal{C}_{\mathcal{N}_{-i, h}}^{n-3}} \frac{\partial^2 f^{r1}}{\partial T_{r1}^2} \frac{\partial^2 f^{r2}}{\partial T_{r2}^2} \dots \frac{\partial^2 f^{r_{n-3}}}{\partial T_{r_{n-3}}^2} \right)}{(-1)^{n+1} \prod_{j \in \mathcal{N}_{-i}} \frac{\partial^2 f^j}{\partial T_j^2} + (-1)^n (n-1) \frac{\partial^2 f^i}{\partial S_i^2} \sum_{r \in \mathcal{C}_{\mathcal{N}_{-i}}^{n-2}} \frac{\partial^2 f^{r1}}{\partial T_{r1}^2} \frac{\partial^2 f^{r2}}{\partial T_{r2}^2} \dots \frac{\partial^2 f^{r_{n-1}}}{\partial T_{r_{n-1}}^2}} \right) \\
& = -\operatorname{sgn} \left( \frac{\partial^2 f^h}{\partial T_h \partial M_h} \right). \quad \square
\end{aligned}$$

This proposition is the analogon to Proposition 3 in Section 2. Whether more resources are invested to punish the rich or the poor depends on the sign of cross-derivative  $\frac{\partial^2 f^h}{\partial T_h \partial M_h}$ . If it is positive, this means that absolute marginal effects of punishments are higher when the victim is poorer. Then, the poor will be treated nastier than the rich.

## 4 Conclusion

Basically, we have derived three main results. The first one is not very surprising. That punishing others can be evolutionarily profitable was quite clear from the beginning. The second is more interesting. Under quite general assumptions it turns out that those who are at the top in the beginning will invest most to stay on the top, or more precisely, to increase their relative advantage. This result does not seem to be in strong contradiction with casual empiricism. Furthermore, it can be tested by analysing experimental data. Fehr and Gächter (1996) conducted an experiment which has a structure similar to our model B. After playing a round of a public good provision game subjects were informed about the outcome and had the opportunity to punish their opponents. Fortunately, Fehr and Gächter also collected sociodemographic data about their subjects—in particular income data. It shows that the amount subjects invest in punishment is positively correlated with their income.<sup>6</sup> Of course, this is a very rough measure, but it illustrates that our result is not without predictive power.

The third result is that harming those who are at the lower tail of the income distribution might be most effective. This result has some flavor of immorality, but it has also some significance in real life. Though modern societies organize a lot of support for the less fortunate, it is still remarkable how often especially poor and weak people are exploited and harmed by others. This may have evolutionary reasons which probably cannot be fully overridden by a process of civilization.

## References

- [1] Abbink, K., G. Bolton, A. Sadrieh, and F.-F. Tang (1996): Adaptive learning versus punishment in ultimatum bargaining, *SFB Discussion Paper*, No. B-381, University of Bonn.
- [2] Akerlof, G.A. (1976): The economics of caste and of the rat race an other woeful tales, *Quarterly Journal of Economics*, 90, 599-617.
- [3] Alchian, A. A. (1950): Uncertainty, Evolution and Economic Theory, *Journal of Political Economy*, 58, 211-221.
- [4] Blume, L. and D. Easley (1992): Evolution and market behavior, *Journal of Economic Theory*, 58, 9-40.
- [5] Blume, L. and D. Easley (1995): Evolution and rationality in competitive markets, *Learning and Rationality in Economics* (eds. A. Kirman and M. Salmon), Oxford/Cambridge.

---

<sup>6</sup>The Pearson correlation coefficient is small (0.22) but highly significant ( $p = 0.002$ ).

- [6] Bolton, G. and R. Zwick (1995): Anonymity versus Punishment in Ultimatum Bargaining, *Games and Economic Behavior*, 10, 95-121.
- [7] De Long, J.B., A. Shleifer, L.H. Summers and R.J. Waldmann (1990): Noise trader risk in financial markets, *Journal of Political Economy*, 98, 703-738.
- [8] Fehr, E. and S. Gächter (1996): Cooperation and punishment in voluntary contribution games, *Working Paper*, University of Zurich.
- [9] Friedman, M. (1953): *Essays in Positive Economics*, Chicago.
- [10] Fudenberg, D. and Tirole, J. (1991): *Game Theory*, The MIT Press.
- [11] Huck, S. and J. Oechssler (1995): The indirect evolutionary approach to explaining fair allocations, *Working Paper*, Humboldt–University.
- [12] Sethi, R., and E. Somanathan (1996): The Evolution of Social Norms in Common Property Resource Use, *American Economic Review*, Vol. 86 No.4, 766-788.



$$|J_{ih}| = -(n-1) \frac{\partial^2 f^i}{\partial S_i \partial M_i} \begin{vmatrix} -\frac{\partial^2 f^1}{\partial T_1^2} & 1 & & & \\ & \ddots & \vdots & & (0) \\ & & 1 & & \\ & & & 1 & \\ (0) & & & & \ddots \\ & & & & 1 & -\frac{\partial^2 f^n}{\partial T_n^2} \end{vmatrix}.$$

Now subtract the row whose main diagonal entry is 1 from all upper (or lower) rows to bring this determinant into an upper (or lower) triangular shape and compute this determinant by simply multiplying the main diagonal entries.  $\square$

**Proof of Remark 3:** Since  $|J'_{ih}|$  is the endogenous-variable Jacobian,  $|J_i|$ , where the  $h$ -th column is replaced by the vector 9 we have

$$|J'_{ih}| = \frac{\partial^2 f^h}{\partial T_h \partial M_h} \begin{vmatrix} n' \frac{\partial^2 f^i}{\partial S_i^2} - \frac{\partial^2 f^1}{\partial T_1^2} & 0 & & & \\ & \vdots & & & \left( n' \frac{\partial^2 f^i}{\partial S_i^2} \right) \\ & & \ddots & 0 & \\ & & & 1 & \\ & & & & 0 \ddots \\ \left( n' \frac{\partial^2 f^i}{\partial S_i^2} \right) & & & & \vdots \\ & & & & 0 & n' \frac{\partial^2 f^i}{\partial S_i^2} - \frac{\partial^2 f^n}{\partial T_n^2} \end{vmatrix}$$

Now proceed again as in the proof of *Remark 1* to get determinants that can easily be computed.  $\square$