Weak moral motivation leads to the decline of voluntary contributions

Charles FIGUIERES, David MASCLET and Marc WILLINGER
February 28, 2009

Abstract
Under the assumption of weakly morally motivated agents, average voluntary contributions decline with repetition of the game, provided that the aggregate moral motivation cannot increase. Our model is compatible with the conditional cooperation hypothesis.

Keywords: Conditional cooperation, voluntary contributions, moral motivation, experiments on public goods games.

JEL:

Acknowledgement 1 Ananish Chaudhuri, Charles Noussair, Dimitry Rtischev, Participants in APESA (Singapore)

1 Introduction

Why are voluntary contributions declining in linear public goods experiments? Does it take time to learn to free-ride? Is there an incentive to build up a cooperative reputation in early rounds? Is there erosion of reciprocity? In contrast to the huge amount of literature concerned with the puzzling fact that subjects over-contribute with respect to their Nash contribution, fewer attempts have been made to explain the decline of the average contribution observed in most linear public goods experiments (see e.g. Isaac et al., 1984, Isaac & Walker, 1988). These attempts boil down to three main explanations: learning, strategic play and reciprocity. While these explanations differ in their expression, they share the common idea that most people try to behave strategically, but are diverted from their goal for various reasons: e.g. they ignore what their best strategy is, they are constrained by some external factors (norms, altruistic reference, ...), or they are uncertain about the behaviour of their partners. Instead, most subjects seem to decide about their current contribution by relying both on some unobservable contribution target and by reacting to observed and/or expected contributions by others.

Our interpretation of the decay is that while most people are morally motivated to contribute to the public good, their moral preferences are weak. By
weak moral motivation we mean that agents have a preferred contribution, but which they might revise depending on their observations (or expectations) of what other people do. Based on the assumption of weakly motivated agents, we propose a model of contribution behaviour that leads to declining average contribution. The central idea is that agents choose their contribution by relying on two dimensions: a "morally ideal contribution" (see Nyborg, 2000, and Brekke et al., 2003) and the observed (or expected) contributions of their group members. The notion of morally ideal contribution is advocated by commitment theories (see e.g. Croson 2007) which rely on Kantian reasoning (Kolm ? Sugden ?). According to these theories, individuals make unconditional commitments to contribute to the public good. The originality of our approach is that we assume that for most people such commitments are weak in the sense that they are sensitive to others' actions. In a contribution context, individuals might therefore be tempted to revise their preferred contribution after observing others' contributions. The extent of such a revision typically varies across individuals: strongly motivated agents will closely stick to their ideal contribution, while weakly motivated agents are prone to revise their morally ideal contribution whenever they observe a gap between own and others' contributions. Our idea of weakly morally motivated agents captures a large spectrum of contribution behaviours: at one extreme, there are unconditional contributors, who always stick to their ideal contribution whatsoever, and at the other extreme there are pure reciprocators who always match the observed group contribution. Note that free-riding behaviour is a particular case of unconditional behaviour, which in a linear public good environment corresponds to always contributing zero.

The assumption of weak moral motivation provides a theoretical basis for imperfect reciprocal behaviour, recently identified in lab experiments (Fischbacher et al., 2001, Fichbacher & Gaechter, 2009, Neugebauer et al., 2009). Imperfect reciprocators are slightly selfish-biases, i.e. they contribute a little less than others' average contribution. Therefore, even in a population of non-selfish agents, the average contribution tends to decrease over time due to the selfish bias. Furthermore, our assumption is compatible both with action-based reciprocity and belief-based reciprocity. By action-based reciprocity we mean that an individual's ideal contribution is sensitive to others' observed contributions, while belief-based reciprocity means that his expectations about others' contributions affects his ideal contribution. For expository reasons, we shall mainly rely on action-based reciprocity.

Let us give a brief sketch of our model. Assume that $N$ players are involved in a repeated game of voluntary contributions to a public good, for which the group contribution is publicly announced after each period. Each player is therefore able to compare his own contribution to the average group contribution. A player who realizes that his contribution is above (below) the average might decrease (increase) his current contribution. We assume that the adjustment with respect to the observed mean contribution is based on updating one's moral motivation. As an illustration, let $x_i^*$ be player $i$'s ex ante ideal contribution and $\bar{x}_t$ the average contribution of the $N$ players in period $t$. Assume that
for $t > 1$ player $i$ revises her ideal contribution according to the linear rule 
\[
\tilde{x}_{it} = \theta_i \tilde{x}_{it-1} + (1 - \theta_i) \bar{x}_{i-1}, \quad \text{with } \theta_i \in [0, 1] \text{ and } \tilde{x}_{i,0} = x_i^*.
\]

The parameter $\theta_i$ can be interpreted as agent $i$’s strength of moral motivation. If $\theta_i = 1$ agent $i$ has as strong motivation and keeps it in the course of the game, whatever the observed contribution of her group. The closer $\theta_i$ to 0, the stronger the revision of her ideal contribution whenever the observed mean differs from her ideal contribution. In this example, a purely reciprocal player is defined by $\theta_i = 0$, and a free-rider by $\theta_i = 1$ and $\tilde{x}_{i,0} = 0$.

While our model is based on action-based reciprocity, it is straightforward to reformulate it for the case of belief-based reciprocity. Rewrite the revision rule as 
\[
\tilde{x}_{it} = \theta_i \tilde{x}_{it-1} + (1 - \theta_i) \bar{x}_{it}, \text{where } \bar{x}_{it} \text{ is agent } i \text{'s expectation of the average contribution of other players for period } t.
\]

The interpretation is now that agent $i$ determines his current contribution by taking into account her expected (current) contribution of other players. If expectations are revised according to the observed contributions by others, i.e. we assume that $\bar{x}_{it} = f(\bar{x}_t)$, with $f'$, we can restrict our model to action-based reciprocity.

Our model is principally designed to account for the decline of the average contribution in linear public goods experiments. However, it is also compatible with most of the stylized facts observed in experiments on voluntary contributions, e.g. significant over-contribution in the end period, high variance of individual contributions and high frequency of individual revision after each period. Furthermore, it accounts for the puzzling restart effect (Andreoni, 1988). Several recent theory papers address a similar issue than ours (see Ambrus & Paterek, 2007, Kandori, 2002 and Klumpp, 2005). Kandori (2002), which is closely linked to our hypothesis, models erosion of norms and morale to explain various phenomena such as the decay of voluntary contributions. (A COMPLETER)

Section 2 provides a brief survey of explanations provided for the decay in average contributions. Section 3 presents the idea of weak moral motivation and studies its implications in a simple linear public good model where agents behave within each period so as to maximize their current payoffs. The simplicity of (linear) utilities and of behaviors (myopia) allows us to focus on the role played by weak moral motivations, but it questions the robustness of the results. Therefore Sections 4 and 5 explore the effect of the weak moral motivations in more complex environments, first when utilities are non linear and then when agents are not myopic. It turns out that the decline of contributions is a somewhat robust consequence of assuming agents who update their moral motivations according to past observations.

## 2 Reasons for declining contributions

Several reasons can explain observed over-contributions in experiments on voluntary contributions to a public good: learning, strategic behavior (i.e. reputation...
or future-oriented behaviour), and social preferences (including altruistic preferences\(^1\) and inequality aversion\(^2\)). As we show, these reasons also explain the decline of average contributions. We restrict our review to the three main hypotheses for accounting for the decline: learning, strategic play and reciprocity (or conditional cooperation).

2.1 Learning

According to the learning hypothesis over-contributions may arise in early rounds because subjects are confused and make errors. As time elapses, and feedback from past rounds becomes available, they realize that they can earn more by over-contributing less, and adjust their current contribution accordingly. Heterogeneous learning leads therefore to the decline in average contributions. Available evidence about the learning hypothesis suggests that learning plays a limited role for the decay. To get a clear picture, it is useful to distinguish learning by introspection and learning by observation of others’ contributions. Learning by introspection has been investigated by Neugebauer et al. (2009) and Houser & Kurzban (2002). Neugebauer et al. (2009) found that repetition without observation has no effect on the average contribution. In their experimental data decay is observed only when information about the average contribution of other group members is provided. Houser & Kurzban (2002) run sessions in which a single human subject is matched with computerized players. The human subject is aware to play against computers and is informed about the current average contribution of the computerized players at the beginning of each round. The authors observe a much sharper decay in sessions with computerized players than in sessions involving only human subjects. Assuming that subjects’ initial confusion is similar across treatments, they estimate that about 50% of the decay is attributable to vanishing confusion. These findings apparently advocate strong introspective learning. However, their two treatments have some critical differences, which make them hardly comparable. In the computerized treatment, the human subject plays a sequential game since he chooses her contribution after observing the others’ average contribution. Recent experiments

\(^1\)One potential explanation relies on the idea that people cooperate because they "take pleasure in others’ pleasure" (see e.g. Dawes and Thaler, 1988). Theory of altruism presented by Andreoni and Miller (1996) assumes that an altruistic player’s utility increases not only in his own payoff but also in the other players’ payoffs. Two forms of altruism are generally given in the literature: “pure altruism” (individuals care about others’ payoffs) and “warm-glow” altruism (individuals enjoy contributing per se). However, both pure and impure altruistic motives cannot adequately describe the decline of contribution observed in public goods experiments. Indeed why would altruistic motives vanish over time?

\(^2\)Another potential explanation relies on inequality aversion. Bolton and Ockenfels (2000) proposed a theory of equity based on the assumption that individuals compare their payoff to that of others and are motivated by their ‘relative’ payoff. Similarly, Fehr and Schmidt (1999)’s theory of inequality aversion postulates that people dislike inequality in payoffs and dislike inequality more if it is to their disadvantage than if it is to their advantage. Note that none of these theories, however, can explain the decline of contribution and the end-game behaviour in repeated games.
showed that sequential contribution environments favor lower contributions by followers (see Fischbacher et al. 2001, Fischbacher & Gächter, 2009, Müller et al., 2008, Mascler et all., 2008). For example, in a two-stage sequential contribution game, Müller et al. (2008) found that "the decline in contributions is much greater and more reliable across the two stages of a given game than across a given stage of successive games". Furthermore, the findings of Houser & Kurzban (2002) leave unexplained the “restart effect” (Andreoni, 1988). It would of interest to know whether subjects react differently to a restart depending on their previous experience with computers or with humans.

Anderson et al. (2004) proposed a model of learning by observation, which accounts for the decay. Their model combines learning direction theory (Selten & Buchta, 1998), i.e. players adjust their current contribution into the direction of increasing payoff - subject to normal errors- and quantal response (McKelvey & Palfrey, 1995). For linear public good games, the model predicts decrease in average contributions, although individual contributions may increase due to random errors. They show that the dynamic process of individual contributions converges to the logit equilibrium distribution of contributions. Noisy adjustment of individual contributions provides therefore a plausible explanation for the decay. However, the model neither explains why contributions are sensitive to the remaining number of periods, as observed in partner treatments, nor does it account for the "restart effect". These effects suggest that both expectations about others’ contributions and the extent of interaction play a crucial role. Furthermore, quantal response is based on the assumption that players are sophisticated enough to be able to best-respond to the stochastic distribution of other players’ contributions. The quantal-response equilibrium is a fixed point where each player’s beliefs about other players’choice distributions matches the equilibrium choice probabilities. Even if quantal response provides a satisfactory way of organizing the data, it is difficult to believe that subjects’ behaviour is compatible with such a model.

While learning probably affects the decay of the average contribution, its relative importance has not yet been clearly established. It is not obvious whether the decay is really faster when a single human subject interacts with computerized players. However, the possibility of observing others’ (average) contribution clearly matters for the decay. Without such observability the average contribution remains constant over time.

2.2 Strategy

The strategy hypothesis is based on the idea that players are forward-looking and choose their current contribution according to their expectations about future interactions. We distinguish between two interpretations of the strategy hypothesis: the "reputation" hypothesis and the "frustrated kind player" hypothesis. According to the "reputation" hypothesis, rational players have an
incentive to make a large contribution in early periods in order to establish a cooperative reputation in their group. The justification is sometimes based on the "crazy player" hypothesis (Kreps et al., 1982), or equivalently on the lack of common knowledge of rationality. If (rational) players believe that there is a crazy player who over-contributes with respect to his Nash contribution, it is rational for them to overcontribute as well, at least in in early periods and switch to their Nash contribution at some later period. There is some experimental evidence about such a behavioural pattern (see e.g. Isaac et al. (1994), Laury (1997), Keser & van Winden (2000)). Also, the fact that decay is slower in longer games (Isaac et al. 1994) seems to support the reputation hypothesis.

Andreoni (1988) offered the first test of the reputation hypothesis by comparing the average contributions of partner groups to stranger groups. Since there is no incentive to develop a cooperative reputation among strangers, one should observe higher over-contributions in partner-groups than in stranger-groups. Andreoni (1988) found exactly the opposite: strangers over-contribute more than partners. Furthermore, he found that the difference increases over time, and complete free-riding is significantly more frequent in the partner treatment. Andreoni’s conclusion is that subjects’ confusion is more likely to be the reason of the decay than strategic play. While his findings seem to destroy the reputation hypothesis as a plausible explanation of the decay, later experiments found mixed evidence about a partner/stranger difference in contributions (see Andreoni & Croson, 2008).

According to the "frustrated kind player" hypothesis (Andreoni, ?), a few cooperative players try to set up a high contribution standard in their group by making large contributions in early periods. By doing so they try to signal their cooperative orientation to their group mates. But since selfish players stick to their Nash contribution, over time cooperators become increasingly frustrated in their attempts to foster cooperation. As the end of the game approaches, frustrated cooperative players stop their trials to induce high contributions, which drives the decay. A crucial ingredient for the decay is heterogeneity of player types plays a crucial for this explanation, since the population must at least contain: unconditional players (e.g. free-riders or warm-glowers) interact with conditional players (see, Fischbacher et al., 2001 and Fischbacher & Gaechter, 2009), who drive the decay process. Experimental evidence shows a high variability of individual contributions. Most subjects revise their contribution after each round (Keser & van Winden, 2000). This finding seems somehow at odds with the hypothesis that there are fixed types.

(TBC)

2.3 Reciprocity

Until recently, reciprocity theories referred to preference heterogeneity for justifying the decay: reciprocal cooperative players are mixed with selfish agents who free-ride on others’ contributions. In a given period, a reciprocal player
who either observes that his contribution is above the average, or who expects others to contribute less, reduces his contribution, favouring thereby the decline of the mean contribution. As we argue below, preference heterogeneity is neither a necessary nor a sufficient condition for the decline in average contributions. Before we address this issue, let us first give some clarification about the notion of reciprocity in the context of voluntary contributions.

The notion of reciprocity, as applied to the provision of public goods, was simultaneously introduced by Kolm (1984) and Sugden (1984). Unfortunately Kolm’s paper was written in French! The interesting idea in Sugden’s paper is that reciprocity is fundamentally defined as a ‘moral’ obligation (1984) of individual i with respect to the group G he belongs to: “Suppose that every member of G except i is making an effort of at least $\xi$ in the production of some public good. Then let i choose the level of effort that he would most prefer that every member of G should make. If this most preferred level of effort is not less than $\xi$, then i is under an obligation to the members of G to make an effort of at least $\xi$”. The model we develop in this paper tries to capture this dimension of morality, to incorporate it in a theory where agent i’s moral obligation can be revised as other group members’ effort become known.

While the moral principle plays a crucial role in Sugden’s definition, later interpretations of reciprocity, especially those that were developed in the light of experimental findings, emphasized the behavioural aspect of the reciprocity principle. In this view, reciprocity is defined as a desire to reward those whose contribution decision signals kindness and to punish those whose contribution decision signals hostility. It relies on the idea that people react to unfair intentions by sacrificing a part of their payoffs in order to punish others, even when there are no reputation gains from doing so (Rabin, 1993, Falk and Fischbacher, 1999, Dufwenberg and Kirchsteiger, 1999). In the game of voluntary contributions to a public good, reciprocal behaviour is often defined as matching others’ contributions (Andreoni, 1995, Palfrey and Prisbey, 1997, Sonnemans et al., 1999, Fischbacher et al., 2001, Kerser and van Winden, 2000 and Brandts and Schram, 2001). Such matching behaviour can be defined in various ways: a reactive reciprocal player will match the previously observed contributions, and a forward looking player will match the expected contributions. If individual contributions are observable reactive players can match the mean, the median, the min (as in Sugden’s theory) or the max (see Croson ??). For the vast majority of public goods experiments, only the average (or total) contribution is observable, which restricts reactive behaviour to mean matching. Most experimental data showed that subjects’ contributions are sensitive to the average contribution of the previous period only. Forward looking reciprocators choose their current contribution according to their expectations about others’ contributions. Note however, than since expectations are likely to be affected by past observed contributions, forward looking reciprocators might respond in a similar way to observed contributions as reactive players do.

Recall that the driving force of the decay, according to reciprocity theories, is preference heterogeneity. Evidence about heterogeneous behaviour in student
subjects samples was mainly provided by Fischbacher et al. (2001), Fischbacher & Gaechter (2009), Burlando & Guala (2005) and Keser & van Winden (2000). Applying the strategy method, Fischbacher et al. (2001) found that about 50% of the subjects are conditional co-operators, 30% are free-riders and 15% are of the “hump-shaped type”, also called “triangle contributors”. Based on a much larger sample (n = 140), Fischbacher and Gächter (2009) found that 55% of their subjects are classified as conditional co-operators, 23% as free-riders and 12% as triangle contributors. Burlando & Guala (2005) crossed different methods to classify subjects. Keser & van Winden (2000) classify subjects according to how they react to the observation of others’ contributions in the previous period. Strong free-riders always stick to their Nash contribution (zero in linear public good games), while weak free-riders contribute zero at least half of the time. Similarly strong cooperators always contribute their whole endowment, while weak co-operators contribute their endowment at least 50% of the time. While these definitions do not allow to classify all subjects, the data of Keser & van Winden (2000) clearly show that the proportion of free-riders (weak or strong) is largest in their stranger treatment while the proportion of cooperators (weak or strong) is largest in their partner treatment.

While player heterogeneity is often taken as a central feature for the decay, it is easy to show that heterogeneity per se is neither a necessary nor a sufficient condition for the decline. To see why, consider a perfect mean matching reciprocator, i.e. a player who matches the average group contribution of the previous period. Assuming that the population is a mixture of perfect reciprocators and unconditional players leads to an average group contribution that can either rise, fall, or stay at a constant level. Let us take an example involving only two players: a perfect reciprocator and an unconditional player who always contributes the same fixed amount in each period. As the game is repeated the contribution of the reciprocal player converges to the fixed contribution of the unconditional player, with a slope that depends on the initial contribution of the reciprocator. The example can be easily extended to any mixed population of a larger size composed of perfect reciprocators and unconditional players. Adding noisy players who contribute a random amount does not guarantee that the mean contribution will either decay or increase. However, if reciprocators are selfishly-biased (Neugebauer et al., 2009) or imperfect (Fischbacher & Gaechter, 2009), the average contribution will always fall. Imperfect, or selfishly-biased, reciprocators are reactive reciprocators who contribute a little less than the observed mean. Evidence about imperfect reciprocity, was provided recently by Neugebauer et al., (2009), Fischbacher & Gaechter (2009) and Fischbacher et al. (2001). Neugbauer et al. (2009), suggest that adaptation of beliefs about others’ contributions combined with selfish bias leads to a downward spiral of contributions. According to Fischbacher & Gaechter (2009) imperfect condi-

---

Except in some extreme and empirically irrelevant cases. For example one could define counter-reciprocal behaviour, which consists in increasing one’s contribution when the mean falls, in order to stop a downward spiral.
tional cooperation is the main driving force behind the decay "Many people's desire to contribute less than others, rather than changing beliefs of what others will contribute over time". Because of such "sticky beliefs" decay can arise in a population of non-selfish agents.

If agents adapt myopically to observed contributions, a small selfish orientation will generate a downward spiral of contributions (Neugebauer et al. 2009). The outcome is more difficult to predict if agents are aware of their impact on others: a few less-selfishly biased agents could over-contribute strategically to curb the decline. However, according to Fischbacher & Gaechter (2009) "Many people's desire to contribute less than others, rather than changing beliefs of what others will contribute over time" generates the decay.

The model we propose captures the selfish bias, and accounts for the fact that such bias can evolve over time. The idea is that an agent’s contribution is the outcome of a trade-off between her material payoff and her moral motivation to contribute to the group, as in Sugden's theory of reciprocity. The selfish orientation becomes stronger as the moral motivation declines with the observation of group contributions.

3 Weak moral motivation in a linear public good model

3.1 What is a weak moral motivation?

Several agents, indexed $i = 1, ..., I$, can contribute voluntarily to a public good. Each individual $i$ has to decide how to split his endowment $w_i$, between his contribution to the public good, $x_i$, and his consumption of the private good, $w_i - x_i$. Using the notation $x_{-i} = \sum_{j \neq i} x_j$, the cardinal representation of agents' preferences with moral motivation are:

$$U^i (x_i, x_{-i}, \hat{x}_i) = w_i - x_i + \beta_i (x_i + x_{-i}) - v_i (x_i - \hat{x}_i) \quad i = 1, ..., n,$$

where $\beta_i \in [0, 1]$ is the marginal utility from consuming the public good $G = x_i + x_{-i}$.

The novel aspect of the above preferences comes from the moral motivation embodied in the functions $v_i(\cdot)$. If $\hat{x}_i$ stands for agent $i$'s moral obligation, then her loss of utility attached to any deviation from her moral obligation is $v_i (x_i - \hat{x}_i)$. This function is assumed to be convex. In addition two natural assumptions about $v_i(\cdot)$ are as follows:

$A_1$: $v_i (0) = 0, \; v_i (x_i - \hat{x}_i) > 0$ iff $x_i \neq \hat{x}_i$.

$A_2$: $v_i' (\cdot) \geq 0 \iff x_i - \hat{x}_i \geq 0$
The first assumption is obvious. The second assumption means that, starting from a situation where agent \( i \) contributes less (more) than her moral obligation, a marginal increase of \( x_i \) reduces (increases) her loss of utility.

We shall conceptualize the moral obligation of each agent as a combination of two logics. There is first an autonomous logic, captured by an ideal level of contribution noted \( x^*_i \geq 0 \), based on a Kantian Categorical Imperative. For instance, \( x^*_i \) could correspond to a Pareto optimal level of contributions. The second logic captures social influences via the average contribution observed in the immediate past, \( \bar{x}_{t-1} \geq 0 \). Finally, the qualified moral obligation is given by a function of the aforementioned variables:

\[
\tilde{x}_{i,t} = \begin{cases} 
  x^*_i, & t = 0, \\
  M^i(x^*_i, \bar{x}_{t-1}), & t = 1, 2, ...
\end{cases}
\]

where function \( M^i(.,.) \) is discontinuous at \( t = 0 \), for there is no previous observations at that date that could be used to qualify the autonomous ethical level.

The moral obligation function satisfies the intuitive properties:

\[
A_3: \frac{\partial M^i}{\partial x^*_i} = M^i_1 \geq 0, \quad \frac{\partial M^i}{\partial \bar{x}_{t-1}} = M^i_2 \geq 0,
\]

\[
A_4: x^{\min}_{it} = \min (x^*_i, \bar{x}_{t-1}) \leq M^i(x^*_i, \bar{x}_{t-1}) \leq x^{\max}_{it} = \max (x^*_i, \bar{x}_{t-1}).
\]

Also, it is assumed that the aggregate qualified moral obligation is bounded above by the aggregate autonomous moral obligation:

\[
A_5: \sum_i M^i(x^*_i, a) \leq \sum_i x^*_i = G^*, \quad \forall a \geq 0.
\]

Because of assumption \((A_4)\), hereafter \( M^i(x^*_i, a) \) is refered to as a weak moral motivation. Regarding Assumption\((A_5)\), note that, while this inequality holds in aggregate, it needs not be true at individual levels, i.e. \( M^i(x^*_i, a) > x^*_i \) for some \( i \) and some \( a \) is a possibility. In addition, this assumption plays an important role to explain the decline of contributions only in so far that the private good consumption and the public good consumption enter linearly in the utility functions, as in (1). In section 5 we study a case of non-linear utility, for which contributions decrease over time, and assumption \((A_5)\) appears as a result.

**Example 2** An illustration of a weak moral motivation function is the following:

\[
\tilde{x}_{i,t} = (1 - \theta_i) x^*_i + \theta_i \bar{x}_{t-1} \ , \quad \theta_i \in [0, 1] \ ,
\]

\[
= x^*_i - \theta_i (x^*_i - \bar{x}_{t-1}) \ .
\]
3.2 Repeated play with myopic contributors

At each period of time, a Myopic Nash Equilibrium (MNE)\(^4\) is a profile of contributions such that each agent’s contribution maximizes his own current utility, given other agents’ contributions:

\[
\max_{x_{i,t}} w_i - x_{i,t} + \beta_i (x_{i,t} + x_{-i,t}) - v_i (x_{i,t} - \bar{x}_{i,t}).
\]

From the first order conditions, interior decisions solve:

\[-1 + \beta_i = \nu'_i(x_{i,t} - \bar{x}_{i,t}), \quad \forall i, \forall t,
\]

thus individual equilibrium contributions at period \(t\) are:

\[x_{i,t} = M^i(x^*_i, \bar{x}_{t-1}) + (v'_i)^{-1}(\beta_i - 1), \quad \forall i.\]

**Proposition 3** At a MNE, the level of public good decreases over time.

**Proof.** The proof is established recursively. Note first that \(G_1 = \sum_i x^*_i + \sum_i (\nu'_i)^{-1}(\beta_i - 1) \leq \sum_i x^*_i\) because \(\sum_i (\nu'_i)^{-1}(\beta_i - 1) \leq 0\). Then observe that \(G_2 = \sum_i M^i(x^*_i, \frac{\bar{G}_1}{t}) + \sum_i (\nu'_i)^{-1}(\beta_i - 1) < G_1\), because \(\sum_i M^i(x^*_i, \frac{\bar{G}_1}{t}) \leq \sum_i x^*_i\) by Assumption A5.

Assume the property \(G_t < G_{t-1}\) holds for \(t = 3, ..., k\), for some \(k\). To complete the proof, it must be established that \(G_{k+1} < G_k\). This is straightforward, for if the property is true until \(t \leq k\), it follows that \(G_{k+1} = \sum_i M^i(x^*_i, \frac{\bar{G}_k}{t}) + \sum_i (\nu'_i)^{-1}(\beta_i - 1)\) is lower than \(G_k = \sum_i M^i(x^*_i, \frac{\bar{G}_k}{t}) + \sum_i (\nu'_i)^{-1}(\beta_i - 1)\), because each \(M^i(., .)\) is an increasing function of its second argument and this argument has fallen, \(G_k < G_{k-1}\).

The dynamics of aggregate contributions \(G_t = \sum_i x_{i,t}\) are:

\[G_t = \sum_i M^i\left(x^*_i, \frac{\bar{G}_{t-1}}{t}\right) + \sum_i (\nu'_i)^{-1}(\beta_i - 1).\]

We now introduce two additional assumptions. The first one stipulates that an increase of the previous level of public good has a less than proportional positive effect on the levels of weak moral motivations:

\[A_6: M^i_2 \leq 1.\]

The second requires that the moral motivation is strong enough to induce a positive level of public good at a MNE even if the previous observable level was zero, i.e. even with an extremely adverse social influence.

\[A_7: \sum_i M^i(x^*_i, 0) \geq - \sum_i (\nu'_i)^{-1}(\beta_i - 1).\]

\(^4\)This concept is not ours. In particular it has been used extensively in the literature on processes (see Drèze and De la Vallée Poussin, 1977, for instance).
Theorem 4 Under Assumptions A5, A6 and A7, the sequence of public good provision converges to a unique positive interior level $G^\infty \in [0; G^*]$. 

Proof. Under Assumption A6, the right hand side of the dynamics (2) is a contraction. Therefore Banach’s fixed point theorem tells us: i) that the dynamics (2) has a unique steady state, ii) the sequence converges towards this steady state. Assumptions A5 and A6 respectively discard the zero and full contributions corner stationary points. ■

Proposition 5 Under Assumption A6, the higher the autonomous ethical level $x_i^*$, the higher the long run level of public good $G^\infty$.

Proof. The long run level of public good solves

$$G^\infty = \sum_i M_i \left( x_i^*, \frac{G^\infty}{T} \right) + \sum_i (v_i')^{-1} (\beta_i - 1).$$

(3)

Using the implicit function theorem:

$$\frac{dG^\infty}{dx_i^*} = \frac{M_i}{1 - \frac{M_i}{T}} > 0$$

under Assumption A6. ■

Proposition 6 The higher the marginal utility of the public good $\beta_i$, the higher the long run level of public good $G^\infty$.

Proof. Using (3) and the implicit function theorem again, one finds:

$$\frac{dG^\infty}{d\beta_i} = \frac{1}{v_i''} > 0,$$

since $v_i(.)$ is a convex function. ■

4 Weak moral motivation in a non linear model

In the linear public good assumption A5 plays a key role for the decay. In this section we show that this condition needs not be imposed in non linear environments. An example is sufficient to show that. Assume agents are identical and their preferences can be represented by the following non linear utility function:

$$U_i (x_{it}, x_{-it}, \tilde{x}_{it}) = \alpha \ln (w - x_{it}) + \beta (x_{it} + x_{-it}) - \gamma \ln \left( v (x_{it} - \tilde{x}_{it}) \right), \quad \alpha, \beta > 0, \gamma \geq 0,$$

$i = 1, ..., n$, where $v (x_i - \tilde{x}_i) = \frac{v}{2} (x_i - \tilde{x}_i)^2$. At a symmetric MNE, individual contributions $x_{it} = x_t \ \forall i$, solve the first order condition:

$$- \frac{\alpha}{w - x_t} + \beta - \frac{2\gamma}{x_t - \tilde{x}_t} = 0.$$  

\footnote{For small values of $\gamma$, the second order condition is satisfied.}
Assume $w\beta > \alpha$. If $\gamma = 0$ (no moral motivation), equilibrium contributions are:

$$x^N = w - \frac{\alpha}{\beta} > 0.$$

If $\gamma > 0$ (moral motivation), the first order condition is a quadratic equation:

$$x^2 - \left(x^N + \bar{x}_t + 2\frac{\gamma}{\beta}\right)x_t + x^N\bar{x}_t + 2\frac{\gamma w}{\beta} = 0.$$

The solution is well-defined if $x^N \leq \bar{x}_t$, and the equation has real roots, i.e. if:

$$\left[x^N + \frac{2\gamma}{\beta}\right]^2 - 4\left(\bar{x}_tx^N + \frac{2\gamma w}{\beta}\right) \geq 0.$$

Assuming this inequality holds for any endogenous value of $\bar{x}_t$, from (??) one observes immediately that both the sum of the roots of this equation and their product is positive, hence both roots are positive: $0 < \bar{x}_t \leq \bar{x}_t$. Further analyzing the roots, it turns out that:

1. $x_t, \bar{x}_t \in ]x^N, \bar{x}_t[, \text{ and,}$
2. $x_t$ is a decreasing funtion of $\bar{x}_t$ whereas $\bar{x}_t$ is an increasing function of $\bar{x}_t$.

This last property, which is crucial for our purpose, is easy to establish. Applying the implicit function theorem to (??), one can check that:

$$\frac{dx}{d\bar{x}_t} |_{x_t} = \frac{x_t}{2x_t - \left(x^N + \bar{x}_t + 2\frac{\gamma}{\beta}\right)} = \frac{x_t}{\bar{x}_t - x_t} < 0, \quad (4)$$

$$\frac{dx}{d\bar{x}_t} |_{\bar{x}_t} = \frac{x_t}{2x_t - \left(x^N + \bar{x}_t + 2\frac{\gamma}{\beta}\right)} = \frac{x_t}{x_t - \bar{x}_t} > 0. \quad (5)$$

Given all those pieces of information, let us now see the dynamics of $\bar{x}_t$, the high MNE. Given an initial motivation $\bar{x}_0 = x^*_t$, a first high MNE settles at $\bar{x}_0 \in ]x^N, \bar{x}_0[$, with a level of public good $G_0 = n\bar{x}_0$. At period $t = 1$, the new moral motivation is $\bar{x}_1 = M^f(x^*_t, \bar{x}_0) < x^*_t$. The new moral motivation falls short of the intrinsic motivation, individually and in aggregate, but this is now a result and no longer an assumption. Then, from (5), the new equilibrium is such that $\bar{x}_1 < \bar{x}_0$, and $G_1 = n\bar{x}_1 < G_0 = n\bar{x}_0$. At period $t = 2$, the new moral motivation is $\bar{x}_2 = M^f(x^*_t, \bar{x}_1) < \bar{x}_1$, implying $\bar{x}_2 < \bar{x}_1 < \bar{x}_0$, and so on and so forth...

By a similar logic the sequence of low MNE, $\bar{x}_t$, increases over repetitions. Overall, if our theoretical explanation is of any relevance as regards the decline of contributions, it means that reals subjects coordinates on the series of high MNE.

Below is a numerical illustration. The parameters are:

$$w = 5, \, \gamma = 0.01, \, \beta = 2/3, \, \alpha = 1, \, \bar{x}_0 = x^* = 4, \, \theta = 0.5.$$
The MNE and the weak moral motivation are reported in the table below, for the first four periods:

<table>
<thead>
<tr>
<th>t</th>
<th>( x_t )</th>
<th>( \bar{x}_t )</th>
<th>( \hat{x}_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.606 2</td>
<td>3.923 8</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3.621 5</td>
<td>3.870 4</td>
<td>3.961 9</td>
</tr>
<tr>
<td>2</td>
<td>3.637 2</td>
<td>3.828 4</td>
<td>3.935 2</td>
</tr>
<tr>
<td>3</td>
<td>3.656 5</td>
<td>3.787 5</td>
<td>3.914</td>
</tr>
</tbody>
</table>

## 5 Non myopic agents

Let \( T \) be the number periods during which agents interact and \( 0 < \sigma \leq 1 \) their discount factor. Their intertemporal utilities are:

\[
\sum_{t=0}^{T} \sigma^t U^i(x_{it}, x_{-it}, \hat{x}_{it}) \, , \, i = 1, \ldots, n,
\]

where

- \( U^i(x_{it}, x_{-it}, \hat{x}_{it}) = w_i x_{it} - \beta (x_{it} + x_{-it} - v_i (x_{it} - \hat{x}_{it})) \), where \( v_i (x_{it} - \hat{x}_{it}) = \frac{v}{2} (x_{it} - \hat{x}_{it})^2 \), \( v_i \) a positive scalar,
- \( \hat{x}_{it} = (1 - \theta_i) x_i^* + \theta_i \bar{x}_{t-1} = M^i(x_i^*, \bar{x}_{t-1}), \quad \theta_i \in [0, 1] \).

We will consider a Markov perfect equilibrium (MPE) for this dynamic public good game. Reasoning backward, in the last period agent \( i \) takes as given \( x_{iT} \) and solves:

\[
\max_{x_{iT}} w_i - x_{iT} + \beta_i (x_{iT} + x_{-iT}) - v_i (x_{iT} - M^i(x_i^*, \bar{x}_{T-1})).
\]

He has a dominant strategy, configured by the average contribution inherited from the previous period:

\[
x_{iT} = \frac{-(1-\beta)}{v_i} + M^i(x_i^*, \bar{x}_{T-1}) \equiv g_{iT}(\bar{x}_{T-1}).
\]

It is worth noting that

\[
g_{iT} = M^i_2(\ldots) = \theta_i. \quad (6)
\]

Those equilibrium strategies can be plugged back into the last period utility, giving each agent’s value function for the last period:

\[
V^i_T(\bar{x}_{T-1}) = U^i(g_{iT}(\bar{x}_{T-1}), \sum_{j \neq i} g_{jT}(\bar{x}_{T-1}), M^i(x_i^*, \bar{x}_{T-1})).
\]

Moving backward to the before last period, each agent’s decision solves:

\[
\max_{x_{i, T-1}} \left\{ w_i - x_{iT-1} + \beta_i (x_{iT-1} + x_{-iT-1}) - v_i (x_{iT-1} - M^i(x_i^*, \bar{x}_{T-2})) \right. \\
+ \sigma V^i_{T-1}\left(\frac{x_{iT-1}+x_{-iT-1}}{n}\right)
\]

The optimal \( x_{iT-1} \) cancels out the addition of several marginal effects:
\(i\) as when agents are myopic, on the current period:

\[-1 + \beta_i - v_i'(x_{iT-1} - M^i(x_i^*, \bar{x}_{T-2})) ,\]

\(ii\) but unlike the case of myopic agents, there is also a marginal effect on the next period \(\sigma \frac{\partial}{\partial x_{iT-1}} V^i_T(\bar{x}_{T-1})\). This expression can be developed as follows:

\[
\sigma \frac{\partial}{\partial x_{iT-1}} V^i_T(\bar{x}_{T-1}) = \sigma \left[ -g'_i + \beta_i \left( g'_i + \sum_{j \neq i} g'_j \right) \right] \frac{\partial \bar{x}_{T-1}}{\partial x_{iT-1}} ,
\]

\[
= \sigma \left[ -g'_i + \beta_i \left( g'_i + \sum_{j \neq i} g'_j \right) \right] \frac{1}{n} ,
\]

\[
= \frac{\sigma}{n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] .
\]

This second marginal effect explains the difference between the myopic and foresighted behaviors. Note that this difference owes nothing to the next period deviation from the moral motivation. Indeed, a marginal increase of \(x_{iT-1}\) has an impact on the next period moral motivation equal to 

\[
\sigma \frac{\partial}{\partial x_{iT-1}} V^i_T(\bar{x}_{T-1}) = \sigma \left[ -g'_i + \beta_i \left( g'_i + \sum_{j \neq i} g'_j \right) \right] \frac{\partial \bar{x}_{T-1}}{\partial x_{iT-1}} ,
\]

\[
\]

\[
= \sigma \left[ -g'_i + \beta_i \left( g'_i + \sum_{j \neq i} g'_j \right) \right] \frac{1}{n} ,
\]

\[
= \frac{\sigma}{n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] .
\]

Overall, the first order conditions are:

\[-1 + \beta - v_i'(x_{iT-1} - M^i(x_i^*, \bar{x}_{T-2})) + \frac{\sigma}{n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] = 0 .
\]

Hence, the best (dominant) response for each agent is:

\[x_{iT-1} = - \frac{(1 - \beta)}{v_i} + \frac{\sigma}{v_in} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] + M^i(x_i^*, \bar{x}_{T-2}) ,
\]

\[= g_{iT-1}(\bar{x}_{T-2}) .
\]

Observe again that:

\[g'_{iT-1} = M^i_{2} (., .) = \theta_i .
\]
Each agent’s value function for the before last period is then:

\[ V_{T-1}^i (\pi_{T-2}) \equiv U^i \left( \sum_{j \neq i} g_{jT-1} (\pi_{T-2}), \sum_{j \neq i} g_{jT-1} (\pi_{T-2}), M^i (x_i^*, \pi_{T-2}) \right) \]

\[ + \sigma V_T^i \left( \frac{\sum_{h=1}^n g_{hT-1} (\pi_{T-2})}{n} \right) \]

Recursively it is possible to construct the agents’ value functions for each date. There is no conceptual difficulty in this exercise but it is tedious and relegated to the Appendix. The important piece of information is that individual problems at date \( t \) generically reads as:

\[
\max_{x_{iT-t}} \left\{ w_i - x_{iT-t} + \beta_i (x_{iT-t} + x_{-iT-t}) - v_i (x_{iT-t} - M^i (x_i^*, \pi_{T-1})) \right. \\
\left. + \sigma V_{T-t+1}^i \left( \frac{x_{iT-t} + x_{-iT-t}}{n} \right) \right\}
\]

And, using the notation \( \Lambda = \frac{\sigma}{n} \sum_{h=1}^n \theta_h = \sigma \bar{\theta} < 1 \), the dominant strategy \( t \) periods before the last can be written generically:

\[
x_{iT-t} = \frac{(1 - \beta)}{v_i} + \sigma \frac{\sum_{j \neq i} \theta_j}{v_i n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=0}^{T-1} \Lambda^h + M^i (x_i^*, \pi_{T-1-t}) \\
\equiv g_{iT-t} (\pi_{T-t-1}) .
\]

In particular, at the first period equilibrium decisions are:

\[
x_{i0} = \frac{(1 - \beta)}{v_i} + \sigma \frac{\sum_{j \neq i} \theta_j}{v_i n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=0}^{T-1} \Lambda^h + x_i^* ,
\]

\[
= g_{i0} (x_i^*) .
\]

We are now in position to investigate whether those contributions could decline over time. Assume that:

\( A_8 \):

\[
- \frac{(1 - \beta)}{v_i} + \sigma \frac{\sum_{j \neq i} \theta_j}{v_i n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=0}^{T-1} \Lambda^h \leq 0 ,
\]

\[
- \frac{(1 - \beta)}{v_i} + \sigma \frac{\sum_{j \neq i} \theta_j}{v_i n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \geq 0 .
\]

Those two conditions are met for instance when agents value sufficiently the public good (\( \beta \) is large enough) and discount heavily the future (\( \sigma \) small enough). We can then establish:
Theorem 7 Under assumption $A_8$, the MPE is characterized by non-decreasing contributions over time.

Proof. Observe first that $x_{i0} \leq x^*_i \forall i$, by the first inequality in Assumption $A_8$, hence $M^i (x^*_i, \pi_0) \leq x^*_i \forall i$. Then, we also have:

$$x_{i0} - x_{i1} = \sigma \left[ \frac{1}{v_{i0}n} \left( -1 + \beta \right) \theta_i + \beta \sum_{j \neq i} \theta_j \right] A^{T-1} + \left( x^*_i - M^i (x^*_i, \pi_0) \right) \geq 0, \forall i,$$

since, by the second inequality in $A_8$ the first term in the right hand side of the above expression is positive and, as seen above $x^*_i - M^i (x^*_i, \pi_0) \geq 0$. Repeating the comparison of successive contributions, one immediately sees that $x_{it}$ is non-increasing over time. ■

When the horizon growths large ($T \to \infty$), this first equilibrium decisions tend to:

$$x_{i0} = -\frac{(1 - \beta)}{v_i} + \frac{\sigma}{v_{i0}n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \frac{1}{1 - \Lambda} + x^*_i.$$

Similarly, considering that period $t$ is the first one, and letting the time horizon go to infinity, it is easy to see that the dominant strategies become stationary feedback rules:

$$x_{it} = -\frac{(1 - \beta)}{v_i} + \frac{\sigma}{v_{i0}n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \frac{1}{1 - \Lambda} + M^i (x^*_i, \pi_{t-1}), (8)$$

Clearly, under Assumption $A_8$, the property of non-increasing contributions carries over to the case of an infinite horizon.

6 Conclusion

TBC

Appendix

A Derivation of the Markov Perfect Equilibrium

In the text, equilibrium decisions for periods $T$ and $T - 1$ have been given. Moving backward to period $T - 2$, each agent’s decision solves:

$$\max_{x_{iT-2}} \left\{ w_i - x_{iT-2} + \beta_i (x_{iT-2} + x_{iT-2} - v_i (x_{iT-2} - M^i (x^*_i, \pi_{T-3})) + \sigma V_{T-1}^i (x_{iT-2} + x_{iT-2}) \right\}. $$

The marginal effects of changing $x_{iT-2}$ are now as follows:
i) As before there are effects on the current utility:

$$-1 + \beta_i - u_i' (x_{iT-2} - M^i (x^*_i, \mathbf{x}_{T-3})),$$

ii) there are also marginal effects on the discounted indirect utility of period $T - 1$:

$$\sigma \left[ -g'_{iT-1} + \beta_i \left( g'_{iT-1} + \sum_{j \neq i} g'_{jT-1} \right) - v_i' (x_{iT-1} - M^i (x^*_i, \mathbf{x}_{T-2})) \right] \frac{\partial \mathcal{U}_{T-2}}{\partial x_{iT-2}},$$

$$= \frac{\sigma}{n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right],$$

iii) and finally, there are marginal effects on discounted indirect utility of period $T$:

$$\sigma^2 \left[ -g'_{iT} + \beta_i \left( g'_{iT} + \sum_{j \neq i} g'_{jT} \right) - v_i' (x_{iT} - M^i (x^*_i, \mathbf{x}_{T-2})) \right] \frac{\partial \mathcal{U}_{T-1}}{\partial x_{iT-2}},$$

$$= \sigma^2 \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=1}^n \frac{g'_{hT-1}}{n} \frac{\partial \mathcal{U}_{T-2}}{\partial x_{iT-2}},$$

$$= \frac{\sigma^2}{n^2} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=1}^n \theta_h \frac{\partial \mathcal{U}_{T-2}}{\partial x_{iT-2}},$$

The first order condition is therefore:

$$-1 + \beta_i - u_i' (x_{iT-2} - M^i (x^*_i, \mathbf{x}_{T-3})) + \frac{\sigma}{n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \left( 1 + \frac{\sigma}{n} \sum_{h=1}^n \theta_h \right) = 0.$$

And the dominant response can be expressed again as:

$$x_{iT-2} = -\frac{(1 - \beta)}{v_i} + \frac{\sigma}{vn} \left[ (-1 + \beta) \frac{\theta_i}{n} + \beta \sum_{j \neq i} \frac{\theta_j}{n} \right] \left( 1 + \frac{\sigma}{n} \sum_{h=1}^n \theta_h \right) + M^i (x^*_i, \mathbf{x}_{T-3}),$$

$$\equiv g_{iT-2} (\mathbf{x}_{T-3}).$$

The marginal effects of changing $x_{iT-3}$ are now as follows:

i) the effects on the current utility are:

$$-1 + \beta_i - u_i' (x_{iT-3} - M^i (x^*_i, \mathbf{x}_{T-4})).$$
\( \text{ii} \) there are also marginal effects on the discounted indirect utility of period \( T - 2 \):

\[
\sigma \left[ -g'_{iT-2} + \beta_i \left( g'_{iT-2} + \sum_{j \neq i} g'_j T-2 \right) - v'_i \left( x_{iT-2} - M^i (x^*_i, \bar{\pi}_{T-3}) \right) \left( g'_{iT-2} - M^i_2 (x^*_i, \bar{\pi}_{T-3}) \right) \right] \frac{\partial \bar{\pi}_{T-3}}{\partial x_{iT-3}},
\]

\[
= \frac{\sigma}{n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right],
\]

\( \text{iii} \) and also, there are marginal effects on discounted indirect utility of period \( T - 1 \):

\[
\sigma^2 \left[ -g'_{iT-1} + \beta_i \left( g'_{iT-1} + \sum_{j \neq i} g'_j T-1 \right) - v'_i \left( x_{iT-1} - M^i (x^*_i, \bar{\pi}_{T-2}) \right) \left( g'_{iT-1} - M^i_2 (x^*_i, \bar{\pi}_{T-2}) \right) \right] \frac{\partial \bar{\pi}_{T-2}}{\partial x_{iT-3}},
\]

\[
= \sigma^2 \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \frac{\sum_{h=1}^{n} g_{h T-2} n}{n} \times \frac{\partial \bar{\pi}_{T-3}}{\partial x_{iT-3}},
\]

\[
= \frac{\sigma^2}{n^2} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=1}^{n} \theta_h.
\]

\( \text{iv} \) and finally, there are marginal effects on discounted indirect utility of period \( T \):

\[
\sigma^3 \left[ -g'_{iT} + \beta_i \left( g'_{iT} + \sum_{j \neq i} g'_j T \right) - v'_i \left( x_{iT} - M^i (x^*_i, \bar{\pi}_{T-1}) \right) \left( g'_{iT} - M^i_2 (x^*_i, \bar{\pi}_{T}) \right) \right] \frac{\partial \bar{\pi}_{T-1}}{\partial x_{iT-3}},
\]

\[
= \sigma^2 \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \frac{\sum_{h=1}^{n} g_{h T-3} n}{n} \times \frac{\sum_{h=1}^{n} g_{h T-2} n}{n} \times \frac{\partial \bar{\pi}_{T-3}}{\partial x_{iT-3}},
\]

\[
= \frac{\sigma^2}{n^3} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \left( \sum_{h=1}^{n} \theta_h \right)^2.
\]

The first order condition is therefore:

\[
-1 + \beta_i - v'_i \left( x_{iT-3} - M^i (x^*_i, \bar{\pi}_{T-4}) \right) + \frac{\sigma}{n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \left[ 1 + \frac{\sigma}{n} \sum_{h=1}^{n} \theta_h + \frac{\sigma^2}{n^2} \left( \sum_{h=1}^{n} \theta_h \right)^2 \right] = 0.
\]

And the dominant response can be expressed again as:

\[
x_{iT-3} = \frac{1}{1 - \beta_i} \left( \sigma \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \left[ 1 + \frac{\sigma}{n} \sum_{h=1}^{n} \theta_h + \frac{\sigma^2}{n^2} \left( \sum_{h=1}^{n} \theta_h \right)^2 \right] + M^i (x^*_i, \bar{\pi}_{T-4}) \right) \equiv g_{iT-2} (\bar{\pi}_{T-3}).
\]
Repeating the logic, and using the notation $\Lambda = \frac{\sigma}{n} \sum_{h=1}^{n} \theta_h$, the dominant strategy $t$ periods before the last can be written generically:

\[
x_{iT-t} = -\frac{(1 - \beta)}{v_i} + \frac{\sigma}{v_i n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \left[ 1 + \Lambda + \Lambda^2 + \ldots + \Lambda^{t-1} \right] + M^i (x^*_i, x_{T-t-1}) ,
\]

\[
= -\frac{(1 - \beta)}{v_i} + \frac{\sigma}{v_i n} \left[ (-1 + \beta) \theta_i + \beta \sum_{j \neq i} \theta_j \right] \sum_{h=0}^{t-1} \Lambda^h + M^i (x^*_i, x_{T-t-1}) ,
\]

\[
\equiv g_{T-t} (x_{T-t-1}) .
\]

References


