Implementation in the K-Level Thinking Environment*

Olga Gorelkina

May 6, 2012

Abstract

Standard mechanism design theory mostly relies on Nash equilibrium concepts. However studies of experimental games suggest that Nash Equilibria are rarely played and provide evidence that subjects may be thinking finite number of iterations. The purpose is to find out whether the standard expected externality mechanism (Arrow, D'Aspermont, Gerard-Varet) retains its properties under iterational thinking. The optimal strategies of finitely-rational players generally deviate from Bayesian Nash Equilibria, though the latter are often good approximation of the outcomes of iterative thinking.

1 Introduction

The idea of relaxing the pervasive common knowledge assumption, often referred to as the Wilson's doctrine, has motivated some recent research in mechanism design. Significant progress has been made in studying implementation in frameworks approaching the universal type space - where higher-order beliefs are virtually unrestricted - as opposed to the naive type spaces (common knowledge), in which mechanisms have been studied previsouly. Bergemann and Morris (2005), to whom much of the progress is due, call attention to the research gap between the classical naive-type setup and the universal type space, the most far-reaching generalization. Indeed there is little we know about functioning of mechanisms in belief frameworks that "realistically" deviate

*I would like to thank Vincent Crawford, Francoise Forges, Thomas Mariotti, David Martimort, Benny Moldovanu, Thomas Rieck for their insightful comments.

from the naive type space. This paper takes advantage of the results of experimental research to identify the relevant belief structure and studies the classical expected externality mechanism (Arrow, D'Aspermont, Gerard-Varet) in that framework, thus filling part of the gap pointed out by Bergemann and Morris.

The experimental reseach in game theory has identified some of the reasoning patterns that persist in human subjects and result in the frequent observations of non-equilibrium outcomes. The *iterative thinking environment*, also referred to as *K-level* model, has proven to be a good fit for the observed behavior. In this model, the agent's type is augmented, with respect to the naive type space, by a cognitive characteristic which puts a restriction on the agent's beliefs. Start with a type space that includes the following elements: payoff type, cognitive ability to carry out a certain number of sequential operations, belief about other people's cognitive abilities, belief about their beliefs and so on. For identification, assume the common prior on *payoff* types. Next, narrow the above type space by imposing the restriction introduced in Nagel (1995): Assume that someone whose ability is to make k iterations believes that any other agent's ability amounts to $k-1$ with probability 1. The formation of beliefs is common knowledge, that is, someone who thinks k iterations is known to have a belief that others think $k-1$, who believe, in turn, that everybody else's ability is $k-2$ and so on. Note that we need to treat the agents who make 1 iteration (1-order agents) separately: their belief is that others play randomly. We assume 1-order agents have a common prior on random strategies, and their prior itself is common knowledge. The latter assumption guarantees that all players of the same level of rationality will have the same mapping from type to strategy. We can therefore merge two elements of the initial type space - the cognitive ability and the beliefs - into a single element, which we label the k-order or k-level of rationality. Now each agent will be fully characterized by his payoff type and level k.

The iterative thinking model may not appeal to an equilibrium-tuned mind, primarily because of the belief inconsistency: Whatever degrees of rationality a set of agents have, their beliefs never reflect the true state. I believe that for real-world one-time interactions this model is nevertheless much more relevant than equilibrium play. For example, in an experimental guessing game, described in Nagel (1995), subjects played strategies quite far from the equilibrium one, and the implied beliefs were inconsistent with the reality; futher experiments show that Nash equilibria are rarely played (e.g. Stahl, Wilson (1994), Kubler, Weizsacker (2004)). Besides, even the classics of economic theory adopts models of behavior with inconsistent beliefs, such as the Stackelberg model.

Recent research develops the theory by looking at the consequences of relaxing common knowledge assumptions, in particular, in the mechanism design problem. In the current paper we also abandon the naive type space and look at individuals' optimal strategies in two games: the expected externality mechanism and the first-price-sealed bid auction. The optimal strategies and the outcomes of these two games are analyzed in the iterative-thinking framework, that is in the assumption that agents' belief system

is the one we described above. The paper proceeds as follows: Section 2 formalizes the framework and provides two examples of the games we are going to study; Section 3 provides a general treatment of the first game, the expected-externality mechanism, and shows that in general the optimal strategies of finite-order rational players deter from truth-telling, but as the order of rationality increases, the optimal strategies converge to truth-telling under certain conditions; Section 4 concludes. Proofs of all propositions and solutions to examples can be found in the Appendix.

2 Sequential Thinking Framework and Motivating Examples

In this section I extend the k-level formalism to games of *incomplete* information. Similar extention has been performed in the Crawford, Iberri, who study the implication of k-level thinking in auctions.

We model sequential thinking deviating from the standard environment in which mechanisms have been studied so far: the common knowledge of rationality. We say that the subject is rational of order $k \in N$ (is a level-k player) if given each of his types he maximizes his expected utility thinking that all his counterparts are rational of order $k-1$. Such behavior induces strategy $s_i^{(k)}(\theta_i)$, $s_i^{(k)} : \Theta_i \to S_i$, where the subscript refers to player i, member of finite set I. Rationality of order 0 implies playing a random strategy. This admits the following representation:

$$s_i^{(0)} \sim \Phi_i : S_i \to [0;1]$$

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} E_{s_i^{(0)}} \left[u_i(s_i, s_{-i}^{(0)}(\theta_{-i}); \theta) \mid \theta_i \right] = \arg\max_{s_i \in S_i} E_{s_i^{(0)}} \left[u_i(s_i, s_{-i}^{(0)}; \theta) \right]$$

$$s_i^{(2)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[u_i(s_i, s_{-i}^{(1)}(\theta_{-i}); \theta) \mid \theta_i \right]$$

...

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[u_i(s_i, s_{-i}^{(k-1)}(\theta_{-i}); \theta) \mid \theta_i \right]$$

where $\theta_i \sim F_i : \Theta_i \to [0;1]$, and $\forall i \in I$ the functions F_i and Φ_i are known.

Thus, for example, a level-2 player assumes that her counterparts behave as if all everyone else was playing randomly. We could say that a profile of strategies $\{s^{(\infty)}(\theta)\}$ constitutes a Bayesian Nash Equilibrium. We impose the following assumptions to

make our analysis more tractable: $S_i = \Theta_i$, *independent types, private values, quasi-linear utilities.* Then in a mechanism implementing social choice rule[1] $K(\theta) : \Theta \to A$ (equivalently $K(s) : S \to A$) individual i who is rational of order k chooses his strategy as follows:

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[v_i(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i) + t_i(s_i, s_{-i}^{(k-1)}(\theta_{-i})) \right],$$

for any $k > 1$ while the rational of order 1 player solves:

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} E_{s_{-i}^{(0)}} \left[v_i(K(s_i, s_{-i}^{(0)}); \theta_i) + t_i(s_i, s_{-i}^{(0)}) \right]$$

We are interested in social choice rules that are efficient, i.e. such that

$$\sum_i v_i(K(\theta_i, \theta_{-i}); \theta_i) \geq \sum_i v_i(\kappa; \theta_i)$$

$for\ \forall \kappa \in A$

In the Bayesian setting efficient social choice rules are (weakly) implemented in the expected externality mechanism (d'Aspremont, Gerard-Varet, 1979, Arrow). In our framework the mechanism induces the following optimal strategies:

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[v_i(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i) + E_{\theta_{-i}} \left[\sum_{j \neq i} v_j(K(s_i, \theta_{-i}); \theta_j) + c \right] \right]$$

for any $k > 1$ and:

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} E_{s_{-i}^{(0)}} \left[v_i(K(s_i, s_{-i}^{(0)}); \theta_i) + E_{\theta_{-i}} \left[\sum_{j \neq i} v_j(K(s_i, \theta_{-i}); \theta_j) + c \right] \right]$$

where c is constant in reported values and represents part of the tranfer intended to balance the buget, i.e. make sure that all the transfers add up to zero. As in the Bayesian setting, we assume that the mechanism designer who tries to implement the efficient rule knows the true distribution of types and makes the players internalize their report's expected effect on the others. Note that in the above equations the expected-externality part of the transfer to agent i, $E_{\theta_{-i}} \sum_{j \neq i} v_j(K(s_i, \theta_{-i}); \theta_j)$, depends only on his report; it assumes that everybody else *will tell* the truth and can be calculated *before* all reports have been made, ensuring the ex post budget balance of the AGV transfers. In the alternative, Clarke-Groves version, would have $\sum_{j \neq i} v_j(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); s_j^{(k-1)}(\theta_j))$ [2]

[1] We assume single-valued social choice rules throughout the paper.

[2] or $\sum_{j \neq i} v_j(K(s_i, s_{-i}^{(0)}); s_j^{(0)})$ for 1-order

instead. The CG tariff is calculated after all reports have been made and it assumes that everybody has told the truth. Recall that the CG mechanism implements in dominant strategy, hence it is optimal to tell the truth at the first and all subsequent orders of rationality. We proceed with the analysis of the standard AGV version.[3]

If order 1 prevails, the players do not expect others to behave optimally, and in particular, they do not expect others to play the equilibrium. Instead, 1-order players have a perception of what reports are likely (or unlikely) to be made - not as functions of random types, but as self-contained random values. Agents who are rational of order 2 know how 1-order players behave and maximize their payoffs correspondingly. If it is the case that optimal response to the mechanism of 1-order players is to tell the truth about their types, then 2-order players would tell the truth too (to maximize the sum of utilities), and so would 3-order players and so on.[4] Thus in a group of any finite order of rationality truth-telling would be restored unless there was no distortion at the very first level. To see what happens if there is distortion, consider a simple example.

Example Consider setting with 2 players and $v_i(\kappa, \theta_i) = \theta_i \kappa - \frac{\kappa^2}{2}$ (Thus efficient rule is linear in the reported types, $K(\theta_1, \theta_2) = \frac{\theta_1 + \theta_2}{2}$)

Solution The optimal strategies for this setting are[5]

$$s_i^{(1)}(\theta_i) = \theta_i + \frac{E\theta_{-i} - Es_{-i}^{(0)}}{2} =$$

$$= \arg\max_{s_i \in S_i} E_{s_{-i}^{(0)}} \left[\theta_i \frac{s_i + s_{-i}^{(0)}}{2} - \frac{(\frac{s_i + s_{-i}^{(0)}}{2})^2}{2} + E_{\theta_{-i}} \left[\theta_{-i} (\frac{s_i + \theta_{-i}}{2}) - \frac{(\frac{s_i + \theta_{-i}}{2})^2}{2} + c \right] \right],$$

$$s_i^{(2)}(\theta_i) = \theta_i - \frac{E\theta_i - Es_i^{(0)}}{4} =$$

[3] The CG transfers are expectationally equivalent to AGV-like transfers, where the mechanism designer "adopts" each player's beliefs when assigns him the transfer (call it AGV-2). In BNE framework, provided that the mechanism designer and the players are rational, both AGV and AGV-2 yield the same outcome - this is why the distinction is not conventional. Note that the K(·) is ex-post implementable in AGV-2 and thus (Bergemann, Morris, 2005) is implementable ex interim in all type spaces. In particular indeed, AGV-2 implements it in the K-level type space.

[4] $s_i^{(2)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[v_i(K(s_i, s_{-i}^{(1)}(\theta_{-i})); \theta_i) + E_{\theta_{-i}} \{ \sum_{j \neq i} v_j(K(s_i, \theta_{-i}); \theta_j) + Const \} \right] =$

$= \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[v_i(K(s_i, \theta_{-i}); \theta_i) + E_{\theta_{-i}} \{ \sum_{j \neq i} v_j(K(s_i, \theta_{-i}); \theta_j) + Const \} \right] = \theta_i$, and so on: $s_i^{(k)}(\theta_i) = \theta_i$

[5] Recall that all players have a common prior on the distribution of payoff types and 1-order agents' prior on random strategies is common knowledge.

$$= \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[\theta_i \frac{s_i + s_{-i}^{(1)}(\theta_{-i})}{2} - \frac{(\frac{s_i + s_{-i}^{(1)}(\theta_{-i})}{2})^2}{2} + E_{\theta_{-i}} \left[\theta_{-i} (\frac{s_i + \theta_{-i}}{2}) - \frac{(\frac{s_i + \theta_{-i}}{2})^2}{2} + c \right] \right],$$

$$s_i^{(3)}(\theta_i) = \theta_i + \frac{E\theta_{-i} - E s_{-i}^{(0)}}{8} =$$

$$= \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[\theta_i \frac{s_i + s_{-i}^{(2)}(\theta_{-i})}{2} - \frac{(\frac{s_i + s_{-i}^{(2)}(\theta_{-i})}{2})^2}{2} + E_{\theta_{-i}} \{ \theta_{-i} (\frac{s_i + \theta_{-i}}{2}) - \frac{(\frac{s_i + \theta_{-i}}{2})^2}{2} + c \right],$$

(The general form is $s_i^{(k)}(\theta_i) = \theta_i + (1/2)^k \Delta_{-i}$ if k odd, $s_i^{(k)}(\theta_i) = \theta_i - (1/2)^k \Delta_i$ if k even, where $\Delta_i = E\theta_i - E s_i^{(0)}$).

There are two interesting features of the AGV mechanism in this example. First, as the order of rationality goes to infinity the players' strategies converge to truth-telling whatever the distributions of their types and of the random strategies. Second, if each player's distribution of type and random strategy are such that their means are equal then there is truth-telling whatever the order of rationality. In the next section we will generalize the equivalence result to any preferences and define conditions under which 1-order players over- or underreport their types (and thus truth-telling behavior is killed in any finite-order groups).

Now let us see what happens when auction bidders think sequentially, the way defined above. Consider a first-price sealed-bid auction with n bidders who maximize their expected gains from the auction.

$$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[(\theta_i - s_i) I\{ s_j^{(k-1)}(\theta_j) < s_i, \ \forall j \in I \} \right]$$

$$= (\theta_i - s_i) \Pr\{ s_j^{(k-1)}(\theta_j) < s_i, \ \forall j \in I \} = (\theta_i - s_i) \prod_{j \in I \setminus i} F_j(s_j^{(k-1)^{-1}}(s_i))$$

for any $k > 1$ and:

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} E_{s_{-i}^{(0)}} \left[(\theta_i - s_i) I\{ s_j^{(0)} < s_i, \ \forall j \in I \} \right]$$

$$= (\theta_i - s_i) \Pr\{ s_j^{(0)} < s_i, \ \forall j \in I \} = (\theta_i - s_i) \prod_{j \in I \setminus i} \Phi_j(s_j)$$

Again, a 1-order bidder does not perceive other bidders as playing strategically; he neither cares about their valuations nor knows that they are maximizing some expected utilities. The only thing he knows is that he receives a positive utility from buying the good if he pays for this a sufficiently low amount (lower than what we call his valuation); he also has a belief on the distribution of other participants' bids. Consider a simple particular case of a first-price auction setting:

Example Assume n symmetric players ($F_i(\cdot) = F(\cdot), \Phi_i(\cdot) = \Phi(\cdot)$), and the distribution of types and of random strategies are both uniform on $[0, 1]$.

Solution

- $s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} E_{s_{-i}^{(0)}} \left[(\theta_i - s_i) I\{s_j^{(0)} < s_i, \ \forall j \in I\} \right] = \arg\max_{s_i \in S_i} (\theta_i - s_i) s_i^{n-1} = \frac{n-1}{n}\theta_i$
- $s_i^{(2)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}} \left[(\theta_i - s_i) I\{s_i^{(1)}(\theta_i) < s_i, \ \forall j \in I\} \right] = \arg\max_{s_i \in S_i} (\theta_i - s_i) [\frac{n}{n-1} s_i]^{n-1} = \frac{n-1}{n}\theta_i$
- ...
- $s_i^{(k)}(\theta_i) = \frac{n-1}{n}\theta_i, \ \forall k$

In this example with uniform distributions, the bidding strategy of any finite-order players is exactly the same as in the Bayesian-Nash equilibrium. We will see later that in general, however, the way we model bidders' behavior is *not* irrelevant: strategies iteratively thinking bidders do not necessarily coincide with BNE-strategies - we will see this in Section 4. Now that we have formalized the framework we proceed from Example 1 to a general treatment of the expected externality mechanism.

3 Expected Externality Mechanism under Iterative Thinking

3.1 Equivalence and Distortion at 1-order of Rationality

In this subsection we look at the general case of preferences and study the optimal behavior of a 1-order rational player in the expected externality mechanism. Recall that the first order of rationality implies maximization of the expected payoff in a game while treating other players as behaving randomly. In the case of an AGV game the optimal strategy as a function of type will be defined as follows:

$$s_i^{(1)}(\theta_i) = \arg\max_{s_i \in S_i} E_{s_{-i}^{(0)}} \left[v_i(K(s_i, s_{-i}^{(0)}); \theta_i) + E_{\theta_{-i}} \{ \sum_{j \neq i} v_j(K(s_i, \theta_{-i}); \theta_j) + \xi_i(s_{-i}^{(0)}) \} \right]$$

$$= \arg\max_{s_i \in S_i} E_{s_{-i}^{(0)}} v_i(K(s_i, s_{-i}^{(0)}); \theta_i) + E_{\theta_{-i}} \sum_{j \neq i} v_{-i}(K(s_i, \theta_{-i}); \theta_j),$$

where $K(s_i, s_{-i})$ is the efficient social choice rule that depends on agents' preferences $v_i(\kappa; \theta_i)$.

In example 1 it was shown that in the sequential thinking environment the optimal behavior in an AGV game may deviate from truth-telling. Here we generalize this result to any (quasi-linear) preferences satisfying the Spence-Mirelees condition, however staying in the 2-player framework. The main contribution of this subsection is the statement of sufficient conditions for lying (under- or overreporting types) as the optimal behavior of a degree-1 player. We find that if order-1 players expect reports to be lower than true types, they will overreport their types, and vice versa. If otherwise they do not have a systematic distinction between types and random strategies, they do not have an incentive to lie. The latter result is stated in the following simple proposition.

Proposition "Equivalence" If the distribution of random strategies and the distribution of types are the same, truth-telling is optimal.

This equivalence claim is easily derived from the definition of 1-order optimal strategy stated in the beginning of the section. The proposition implies that sequential thinking in an AGV game produces the same result as suggested by the equilibrium analysis.

In the rest of the section we look at the non-trivial case, when 1-order players anticipate biased reports from their counterparts. The following proposition gives us the idea of how exactly 1-order player's report is going to be distorted. Here we require the preferences be such that the mechanism's sensitivity to one agent's report is independent of the other agent's report. This sensitivity is formalized in the cross-derivative of the efficient social choice rule and we set it to zero at this point. Further we will abandon this assumption and see how the sensitivity of SCR affects our conclusions. We also assume that the preferences satisfy Spence-Mirelees condition: $\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(\kappa, \theta_i) > 0$ $\forall \kappa \in A, forall \theta_i \in \Theta_i$ ("SMC with positive sign") or $\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(\kappa, \theta_i) < 0$ $\forall \kappa \in A,\ \forall \theta_i \in \Theta_i$ ("SMC with negative sign").

Proposition "Linear SCR" If SMC holds, and $\frac{\partial^2 K}{\partial s_i \partial s_{-i}}(s_i, s_{-i}) = 0$, then for any (interior) types we observe underreporting, if the distribution of random strategies 1st-order stochastically dominates the distribution of types, - and overreporting, if the distribution of types 1st-order stochastically dominates the distribution of random strategies.

The proposition tells us that 1-order rational players misreport their types whenever the distributions of types and of strategies are different in a certain sense. Namely, if player A expects player B to report a higher type than B has on average, then

A will report a lower type than he actually has (and vice versa), even if this will push the choice further from what A desires according to his preferences. What is the intuition behind that? In the AGV mechanism agent A gets utility from the social choice based on his and B's reported preferences plus the expected payoff of agent B had he told the truth to the principal. Suppose first that a high type values the size of the alternative more than a low type ("positive SMC"). If agent A knows that agent B would overreport his preferred alternative on average, then - since A benefits from satisfying B's *true* preferences - he would adjust the social choice downward by underreporting himself. If higher types prefer lower alternatives, then B's overreporting makes the chosen alternative lower and A overreports to shift it back up. In either case 1-order rational player compensates the counterpart's "foolishly" biased behavior by misreporting their types in the opposite direction.

The assumption of zero-sensitivity we make in Proposition "Linear SCR" corresponds to the cases when the social choice function is linear in the reported types - for example, the chosen alternative equal to the arithmetic mean of preferred alternatives (types), as we had in Example 1. In such cases agent A may not bother about second-order effects when he tries to adjust the mechanism's choice with his report. However under more general conditions second-order effects may come into play. Namely, if agent A (of a very high or a very low type) knows that his misreporting affects the mechanism's reaction to B's report in such a way that the total distortion becomes even stronger, he may prefer not to misreport in the direction Proposition "Linear SCR" states. Thus we have to exclude certain type ranges when asserting that there are incentives to lie. The following four propositions state similar results to Proposition "Linear SCR", but only for agents having sufficiently low or sufficiently high types. The cases are broken down into four groups according to two criteria: whether higher types prefer *higher* or *lower* alternatives (respectively, SMC with *positive* or with *negative* sign), and whether the chosen alternative's increment due to an increase in one agent's report *increases* or *decreases* with the other agent's report (respectively, *positive* or *negative* cross-derivative). In the four following propositions we impose an additional technical assumption, the monotone likelihood ratio property (MLRP) that we did not require in Proposition "Linear SCR". MLRP should be understood as corresponding to whether the distribution of types dominates or the distribution of random strategies dominates.

Proposition "+ +" If SMC holds with **positive** sign, $\frac{\partial^2 K}{\partial s_i \partial s_{-i}}(s_i, s_{-i}) \geq 0$, and MLRP holds, then for **sufficiently low** types we observe underreporting, if the distribution of random strategies 1st-order stochastically dominates the distribution of types overreporting, if the distribution of types 1st-order stochastically dominates the distribution of random strategies.

Proposition "+ -" If SMC holds with **positive** sign, $\frac{\partial^2 K}{\partial s_i \partial s_{-i}}(s_i, s_{-i}) \leq 0$, and MLRP holds, then for **sufficiently high** types we observe underreporting, if the distribution of random strategies 1st-order stochastically dominates the distribution of

types overreporting, if the distribution of types 1st-order stochastically dominates the distribution of random strategies.

Proposition "- +" If SMC holds with **negative** sign, $\frac{\partial^2 K}{\partial s_i \partial s_{-i}}(s_i, s_{-i}) \geq 0$, and MLRP holds, then for **sufficiently high** types we observe underreporting, if the distribution of random strategies 1st-order stochastically dominates the distribution of types overreporting, if the distribution of types 1st-order stochastically dominates the distribution of random strategies.

Proposition "- -" If SMC holds with **negative** sign, $\frac{\partial^2 K}{\partial s_i \partial s_{-i}}(s_i, s_{-i}) \leq 0$, and MLRP holds, then for **sufficiently low** types we observe underreporting, if the distribution of random strategies 1st-order stochastically dominates the distribution of types overreporting, if the distribution of types 1st-order stochastically dominates the distribution of random strategies.

Proposition "++" tells us that if higher types have higher valuation and the efficient social choice rule is more sensitive to agent 1's reported type if agent 2's report is high, then low-valuation players will tend to lie on their type - so as to adjust the biased reports of their co-players. The result is the same as in Proposition "Linear SCR", except that we no longer assert that high types will necessarily do so. Consider for example agent A whose type is higher than the expected agent B's true or reported type. According to our previous reasoning, agent A would like to overreport if B underreports[6]; but if he does so under the conditions of Proposition "++", the mechanism may become more sensitive to B's underreporting and A's efforts will be in vain (actually the reasoning would be more complicated than that). In all four propositions, we include in the sufficient condition the type ranges that correspond to a sufficiently (for given distributions) weak sensitivity of the social choice rule to the other agent's report. We claim that such types will underreport if $F(t) > \Phi(t)\ \forall t$ and overreport if $\Phi(t) > F(t)\ \forall t$.

We can sum up this subsection by the following. First, if the random strategies are, in the minds of 1-order players, distributed equivalently to how types are distributed, then in the expected externality mechanism 1-order players report their true types. This implies that 2-order players will tell the truth, too, and so will 3-order players etc. Thus sequential thinking yields the same outcome as if agents were playing Bayesian Nash equilibrium, independently of what order of rationality the players as a group have. Second, if distributions are *not* the same, then 1-order players usually misreport their types so as to compensate the biased reports of their co-players. There may be exceptions from this rule if the agent has an extreme type and his "compensating" behavior may affect the social choice in an undesirable way, but such situations ought to be perceived as rare. Looking at the proof of the propositions we conjecture that the results extend to the n-player symmetric framework.

[6]By underreporting we mean here that B's type distribution 1st-order stochastically dominates B's random report distribution.

3.2 Distortion at Higher Orders of Rationality and Convergence

In motivating example 1 we saw that the distortion of reports by 1-order players translates into the optimal strategies of 2-order rational players, 3-order and so on, with the size of the distortion decreasing. The limiting optimal strategy in the example is truth-telling; in this subsection we state a similar result for a somewhat more general setting, by looking at the case of an arbitrary linear social choice rule. Recall that in the discussion of 1-order optimality (the previous subsection) linear SCR case yielded the same behavior as a most general SCR case, but for all possible types. Here we do not go beyond the linear case, leaving it as an approximation for the general case of SCR. We remain in 2-player setting, as before.

Proposition "Convergence" If the social choice rule is linear in reported types, then in the AGV mechanism the expected absolute deviation of reported types from true ones decreases with the order of rationality. The sign of the expected deviation changes each time the order of rationality grows by one.

The first part of the proposition states average convergence to truth-telling in the expected externality mechanism with sequentially thinking players. The second part states that the strategies in any linear-SCR case display the same pattern that we observed in example 1. If 2-order players overstate their type in the game, then 3-order players will understate them. In fact, this is good news for the AGV mechanism: if the group of agents is a mix of, say, 2- and 3-order rational players, then the expected chosen alternative will be closer to the one maximizing the true welfare.

In this section we have shown that agents' expectations of other people's behavior affects the optimal strategies in the d'Aspremont-Gerard-Varet mechanism. Thus if we wish to implement an efficient social choice function we need to use all available information about agents' anticipation of the play of others and adjust the mechanism in such a way that it would account for biases and recover truth-telling behavior. Consider the following illustration. A group of agents is eligible to vote for a transition from status-quo alternative A to alternative B. It is common knowledge that each agent has a preferred alternative and gets a utility of 1 from it. If some voter has not submitted his vote by the deadline, he is considered to prefer the status quo. AGV mechanism is applied, so each voter gets the expected externality of his vote on the others. Suppose every voter of the group puts a 50/50 prior ("has no idea") on both the true preferences and the random votes (he considers the votes random, if, for instance, he does not expect the others to understand the expected externality procedure). Further, everyone supposes that around 20 % of the votes will not be submitted by the deadline due to voters' absence, post delay or other reasons, independently of the choices (potentially) made - thus the ultimate perceived distribution of registered reports (recall, default is

A) will be 60/40. Then the someone who is of order 1 and actually prefers A might submit a vote for B, because the submitted votes, in his eyes, will be biased towards A which will not reflect the group's true preferences, from satisfying which this agent benefits (expected externality mechanism). If the group thinks this way we may end up choosing B, even if it is not actually preferred by the majority. Now if an agent assumes that all others put 50/50 on preferences and random strategies, and expect 20% of the votes to be lost, he might find it profitable to vote for A even if his preferred alternative is B (we require that he himself puts a 50/50 prior on types and expects *all* votes to be submitted in time). If there is a sufficient number of such agents we can choose A even if B is preferred by the majority, and so on. Let us move to the continuous case to apply our result. Suppose that agents are insensitive to small differences in outcomes from strategies A and B and play a mixed strategy when the difference is small. The larger the difference, the more likely they are to play the strategy that yield higher payoff. Then our propositions imply for the voting example that the probability of choosing a wrong alternative decreases as the order of rationality in a group goes up - since each agent reports the truth with higher probability.

4 Conclusion

This paper looks at the functioning of mechanisms, which have been either designed or previously studied under the assumption of common knowledge of rationality, - in a framework that deters from it. Our motivation to apply the new framework to these mechanisms stems in the apparent frailness of the Bayesian -Nash equilibrium in the case of simultaneous-move one-time interactions. In reality, players in a non-repeated game might have no idea that they are "supposed to play" equilibrium strategies, and experiments find sufficient evidence for that. What we propose here as a way the players form their strategies is *iterative thinking*, in which we follow Nagel (1995). An iteratively thinking player optimizes the game's outcome with respect to his strategy, accounting for other players' behavior, who account for all other players' behavior etc., assuming that he can always make 1 iteration more than the others. He treats all other players as having the same capacity in terms of the number of iterations they can make and the ability to optimize; in the simplest case he thinks they do not optimize at all and play randomly. Experiments show that the number of iterations people make in their minds while playing simple simultaneous-move games is between 1 and 2, which prompts us to focus on these levels when we study the mechanisms. We find that in general people's behavior will deviate from the equilibrium prediction, however the conditions under which there is equivalence do not seem to be too demanding.

In the first part of the analysis we looked at how the expected externality mechanism (D'Aspremont, Gerard-Varet) works in the sequential thinking environment. This mechanism was designed so as to induce truthful revelation of types in a particular setting: Bayesian-Nash equilibrium. As BNE relies on the common knowledge of ratio-

nality, it was intriguing to find out under which conditions the mechanism still yields truth-telling behavior as we proceed to an alternative framework. It appears, that if the perceived distribution of random strategies coincides with the perceived distribution of types, then at any number of consecutive iterations we get truth-telling. If the players expect that for some reason their opponents' reports will first-order stochastically dominate the type distribution, they will tend to (certain type ranges necessarily will) underreport their types. We get the symmetric result, too. Further, each subsequent order of rationality yields a smaller absolute value of deviation from truth-telling, and the direction of the deviation switches each time. Thus we observe "compensating" behavior of finite-order players in an AGV game, which implies lying in the opposite direction to the anticipated bias in other agents' report. In the limit, their strategies converge to the equilibrium one, that is, to truth-telling. This convergence result is similar to what is observed in experimental guessing games (e.g. Nagel, 1995): with each repetition of the game players choose strategies closer and closer to equilibrium. Our results were stated for a 2-player framework (conjecturally, they can be extended to n players) and the mechanism that assigns expected-externality transfers basing on each agent's report and the distribution of types of the other players. An alternative version of the mechanism would assign transfers basing on actual reports made. This set-up was found to induce truth-telling for any players who maximize their expected payoff.

What our analysis implies for the use of the expected externality mechanism, is that the way of thinking of subjects has to be taken into account whenever possible. The voting example we looked at in the end of the AGV section prompts that the mechanism may lead to different outcomes in more and less sophisticated groups. Besides, the practical organization of the mechanism would also affect the strategies and outcomes through the agents' beliefs about the registered reports. Hence the rules of the game should be stated *so* clearly and the organizers should show *so much* commitment to them, that all players would not only understand the rules and know their reports would be counted, but also be sure that all *other* players understand the rules and know their reports will be counted. Futhermore, whenever possible, the transfers ought to be computed basing on all agents' reports under the assumption they have told the truth. If the way of computing is made common knowledge, then all optimizing agents will report truthfully.

This paper has shown that if real-world people think finite number of iterations instead of always playing Bayesian-Nash equilibrium, our perceptions of standard mechanisms should change: a first-price auction may no longer be efficient and the expected externality mechanism may yield the choice of "wrong" alternatives. However, in many natural cases BNE still gives correct predictions or, at least, good approximations of what happens in iterative thinking reality. Besides, our convergence results suggest that in repeated interactions, when the agents can partially observe the strategies of others, equilibrium will become an even better approximation.

A Appendix

Proof of Propositions 5 to 10 The first-order conditions of a rational of order 1 player:

$E_{s_{-i}^{(0)}}[\frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})] + E_{\theta_{-i}}[\frac{\partial v_{-i}}{\partial \kappa}(K(s_i, \theta_{-i}); \theta_{-i})\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})] = 0$

Note that $\frac{\partial v_{-i}}{\partial \kappa}(K(s_i, \theta_{-i}); \theta_{-i}) + \frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i) = 0$ (since $K(s_i, s_{-i})$ is efficient)

Thus we can rewrite the second term in the f.o.c.[7]:

$0 = E_{s_{-i}^{(0)}}[\frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})] - E_{\theta_{-i}}[\frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i)\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})]$

$= \int \frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) - \int \frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i)\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})dF(\theta_{-i})$

Take the second term and integrate by part:

$\int \frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i)\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})d\{F(\theta_{-i}) - 1\} =$

where $\underline{\theta_{-i}}$ is the lower bound of the support of $F(\cdot)$. We changed the name of the integration variable to

Modify the first term by taking Taylor expansion (Peano form) under the integral:

$\int \frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) =$

$= \int [\frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(0)}); s_i) + \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)(\theta_i - s_i)]\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})$

where $\widehat{\theta}_i$ is between s_i and θ_i.

$= \int \frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(0)}); s_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})d\{\Phi(s_{-i}^{(0)}) - 1\} + \int \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)(\theta_i - s_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) =$

$= -\frac{\partial v_i}{\partial \kappa}(K(s_i, \underline{s_{-i}^{(0)}}); \theta_i)\frac{\partial K}{\partial s_i}(s_i, \underline{s_{-i}^{(0)}}) - \int \{\Phi(t) - 1\}d\frac{\partial v_i}{\partial \kappa}(K(s_i, t); s_i)\frac{\partial K}{\partial s_i}(s_i, t) +$

$+ \int \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)(\theta_i - s_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})$

where $\underline{s_{-i}^{(0)}}$ is the lower bound of the support of $\Phi(\cdot)$. Assuming that $\underline{s_{-i}^{(0)}} = \underline{\theta_{-i}}$(we further denote it $\underline{t}$) the f.o.c. becomes:

$\int \{F(t) - \Phi(t)\}d\frac{\partial v_i}{\partial \kappa}(K(s_i, t); s_i)\frac{\partial K}{\partial s_i}(s_i, t) +$

$+(\theta_i - s_i)\int \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, s_{-i}^{(0)}); \widehat{\theta}_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})d\Phi(s_{-i}^{(0)}) = 0$

$s_i^{(1)}(\theta_i) - \theta_i \equiv \frac{\int \{F(t) - \Phi(t)\}d\frac{\partial v_i}{\partial \kappa}(K(s_i^{(1)}(\theta_i), t); s_i^{(1)}(\theta_i))\frac{\partial K}{\partial s_i}(s_i^{(1)}(\theta_i), t)}{\int \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i^{(1)}(\theta_i), s_{-i}^{(0)}); \widehat{\theta}_i)\frac{\partial K}{\partial s_i}(s_i^{(1)}(\theta_i), s_{-i}^{(0)})d\Phi(s_{-i}^{(0)})}$

---

[7] s.o.c. $E_{s_{-i}^{(0)}}[\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, s_{-i}^{(0)}); \theta_i)[\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(0)})]^2 + \frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(0)}); \theta_i)\frac{\partial^2 K}{\partial s_i^2}(s_i, s_{-i}^{(0)})] -$
$-E_{\theta_{-i}}[\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \theta_{-i}); s_i)[\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})]^2 + \frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i)\frac{\partial^2 K}{\partial s_i^2}(s_i, \theta_{-i}) + \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \theta_{-i}); s_i)\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})] \mid_{\substack{s_i = \theta_i \\ F(\cdot) = \Phi(\cdot)}} =$
$= -E_{\theta_{-i}}[\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \theta_{-i}); s_i)\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})] < 0$ (see Lemma)

(Note that if $F(t)-\Phi(t)\equiv 0$, there is no distortion, thus Proposition EQUIVALENCE is proven)

Let's look at the signs of expressions. The proof that the denominator is positive is in Lemma 12 (see below). The nominator :

$\int_{\underline{t}}^{+\infty}\{F(t)-\Phi(t)\}d\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\frac{\partial K}{\partial s_i}(s_i,t)=$

$=\int_{\underline{t}}^{+\infty}\{F(t)-\Phi(t)\}[\frac{\partial^2 v_i}{\partial\kappa^2}(K(s_i,t);s_i)\frac{\partial K}{\partial s_{-i}}(s_i,t)\frac{\partial K}{\partial s_i}(s_i,t)+$

$+\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)]dt$

If $\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)=0$ then the term in brackets is negative:

$\frac{\partial^2 v_i}{\partial\kappa^2}(K(s_i,t);s_i)<0$ we assume concavity of preferences in κ (thus s.o.c. of the SCR problem is satisfied)

$\frac{\partial K}{\partial s_{-i}}(s_i,t)\frac{\partial K}{\partial s_i}(s_i,t)>0$ from SMC and Lemma 12

Thus Proposition "Linear SCR" is proven.

Otherwise - any $\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)$ - we need to decompose the denominator. Start with the case of Proposition "++":

$\frac{\partial^2 v_i}{\partial\kappa\partial\theta_i}(\kappa,\theta_i)>0$, $\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)\geq 0$. The denominator:

$\int_{\underline{t}}^{+\infty}\{F(t)-\Phi(t)\}d\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\frac{\partial K}{\partial s_i}(s_i,t)=$

$$=\underbrace{\int_{s_i}^{+\infty}\{F(t)-\Phi(t)\}\left[\underbrace{\frac{\partial^2 v_i}{\partial\kappa^2}(K(s_i,t);s_i)}_{-}\underbrace{\frac{\partial K}{\partial s_{-i}}(s_i,t)}_{+}\underbrace{\frac{\partial K}{\partial s_i}(s_i,t)}_{+}+\underbrace{\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)}_{-}\underbrace{\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)}_{+}\right]dt}_{\text{"first term"}}+$$

$$+\underbrace{\int_{\underline{t}}^{s_i}\{F(t)-\Phi(t)\}d\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\frac{\partial K}{\partial s_i}(s_i,t)}_{\text{"second term"}}$$

It was convenient to break down the integral into 2 parts since $\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)$ decreases in t.[8] and $\frac{\partial v_i}{\partial\kappa}(K(s_i,s_i);s_i)=0$. First term, in brackets:

$\frac{\partial^2 v_i}{\partial\kappa^2}(K(s_i,t);s_i)<0$, concavity of preferences in κ

$\frac{\partial K}{\partial s_{-i}}(s_i,t)>0$, $\frac{\partial K}{\partial s_i}(s_i,t)>0$ from positive SMC and Lemma 12

$\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)$ for $t\leq s_i$ $\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)$ by the assumption. Thus

$$\left[\underbrace{\frac{\partial^2 v_i}{\partial\kappa^2}(K(s_i,t);s_i)}_{-}\underbrace{\frac{\partial K}{\partial s_{-i}}(s_i,t)}_{+}\underbrace{\frac{\partial K}{\partial s_i}(s_i,t)}_{+}+\underbrace{\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)}_{-}\underbrace{\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)}_{+}\right]$$ is negative. The

[8] $\frac{\partial^2 v_i}{\partial\kappa^2}(K(s_i,t);s_i)\frac{\partial K}{\partial t}(s_i,t)<0$

second term:

$$\int_{\underline{t}}^{s_i}\{F(t)-\Phi(t)\}d\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\frac{\partial K}{\partial s_i}(s_i,t)=$$

$$\underbrace{\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)}_{=0}\tfrac{\partial K}{\partial s_i}(s_i,s_i)\{F(s_i)-\Phi(s_i)\}+\tfrac{\partial v_i}{\partial\kappa}(K(s_i,\underline{t});s_i)\tfrac{\partial K}{\partial s_i}(s_i,\underline{t})\underbrace{\{F(\underline{t})-\Phi(\underline{t})\}}_{=0}-$$

$$-\int_{\underline{t}}^{s_i}\{F(t)-\Phi(t)\}\tfrac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\tfrac{\partial K}{\partial s_i}(s_i,t)d\{F(t)-\Phi(t)\}$$

$$=-\int_{\underline{t}}^{s_i}\{F(t)-\Phi(t)\}\tfrac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\tfrac{\partial K}{\partial s_i}(s_i,t)\{f(t)-\varphi(t)\}dt$$

$$\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\geq 0$$

for $t\leq s_i$

$$\frac{\partial K}{\partial s_i}(s_i,t)>0$$

First look at the case $\Phi(\cdot)$ s.d. $F(\cdot)$: $F(t)-\Phi(t)>0\ \forall t\Rightarrow$ the first term is negative. If $f(s_i)-\varphi(s_i)>0$, then the second term is negative, too: The MLRP assumption implies that $\frac{f(t)}{\varphi(t)}$decreases in t; thus under the integral $f(t)-\varphi(t)>0$ and there exists a (unique) t^*such that $f(t^*)-\varphi(t^*)=0$ (see graph). So for θ_i such that $s_i^{(1)}(\theta_i)\leq t^*$the result is established: we have a sufficient (but not necessary[9]) condition for underreporting types in the case $\Phi(\cdot)$ s.d. $F(\cdot)$.

Now suppose that $F(\cdot)$ s.d. $\Phi(\cdot)\Longrightarrow$ the first term is positive. MLRP then implies $\frac{\varphi(t)}{f(t)}$ decreases in t and by the same reasoning for θ_i low enough the second term is positive, too, and we get overreporting of types.

Proposition "++" is now proven.

To prove Proposition "+ -" ($\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)\leq 0$), change the decomposition of the nominator:

$$\int_{\underline{t}}^{+\infty}\{F(t)-\Phi(t)\}d\tfrac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\tfrac{\partial K}{\partial s_i}(s_i,t)=$$

$$=\int_{\underline{t}}^{s_i}\{F(t)-\Phi(t)\}[\underbrace{\frac{\partial^2 v_i}{\partial\kappa^2}(K(s_i,t);s_i)}_{-}\underbrace{\frac{\partial K}{\partial s_{-i}}(s_i,t)}_{+}\underbrace{\frac{\partial K}{\partial s_i}(s_i,t)}_{+}+$$

$$+\underbrace{\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)}_{+}\underbrace{\frac{\partial^2 K}{\partial s_i\partial s_{-i}}(s_i,t)}_{-}]dt+$$

$$+\int_{s_i}^{+\infty}\{F(t)-\Phi(t)\}d\tfrac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\tfrac{\partial K}{\partial s_i}(s_i,t)$$

Given that $\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)$ decreases in t, we have that for $t\leq s_i$ $\frac{\partial v_i}{\partial\kappa}(K(s_i,t);s_i)\geq 0$ and thus the term in brackets is again negative. Integrating the second term by part

[9]The integral may remain negative for some time as s_i goes beyond t^*.

we obtain $-\int_{s_i}^{+\infty} \underbrace{\frac{\partial v_i}{\partial \kappa}(K(s_i,t);s_i)}_{-}\underbrace{\frac{\partial K}{\partial s_i}(s_i,t)}_{+}\{f(t)-\varphi(t)\}dt$. Since we want both parts of the nominator to have the same sign and MLRP applies, we need to have s_i sufficiently *high* now (or θ_i such that $s_i^{(1)}(\theta_i) \geq t^*$). Proposition "+-" proven.

The proofs of Propositions "- +" and "- -" that assume SMC with negative sign are similar to the presented above. With negative SMC $\frac{\partial K}{\partial s_i}(s_i, s_{-i})$ becomes *negative* and $\frac{\partial v_i}{\partial \kappa}(K(s_i,t);s_i)$ *increases* in t.

Lemma Under SMC $\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i^{(1)}(\theta_i), s_{-i}^{(0)}); \widehat{\theta}_i)\frac{\partial K}{\partial s_i}(s_i^{(1)}(\theta_i), s_{-i}^{(0)}) > 0$ for any $\theta_i, \widehat{\theta}_i, s_{-i}^{(0)}$.

Proof $\frac{\partial v_i}{\partial \kappa}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta}_i) + \frac{\partial v_{-i}}{\partial \kappa}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta_{-i}}) \equiv 0$ (f.o.c. for finding the efficient rule)

Differentiate w.r.t. $\widetilde{\theta}_i$:
$\frac{\partial K}{\partial s_i}(\widetilde{\theta}_i, \widetilde{\theta_{-i}})[\frac{\partial^2 v_i}{\partial \kappa^2}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta}_i) + \frac{\partial^2 v_{-i}}{\partial \kappa^2}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta_{-i}})] + \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta}_i) = 0$
From s.o.c. of the same problem, $\frac{\partial^2 v_i}{\partial \kappa^2}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta}_i) + \frac{\partial^2 v_{-i}}{\partial \kappa^2}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta_{-i}}) < 0$

Thus, $sgn(\frac{\partial K}{\partial s_i}(\widetilde{\theta}_i, \widetilde{\theta_{-i}})) = sgn(\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(\widetilde{\theta}_i, \widetilde{\theta_{-i}}), \widetilde{\theta}_i))$. Substitute $\widetilde{\theta}_i$ by $s_i^{(1)}(\theta_i)$, $\widetilde{\theta_{-i}}$ by $s_{-i}^{(0)}$ and get $sgn(\frac{\partial K}{\partial s_i}(s_i^{(1)}(\theta_i), s_{-i}^{(0)})) = sgn(\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i^{(1)}(\theta_i), s_{-i}^{(0)}), s_i^{(1)}(\theta_i)))$. With SMC (sign of $\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(\kappa, \theta_i)$ is the same for all (κ, θ_i)) the result is proven.

Proof of Proposition 11 f.o.c. for finding $s_i^{(k)}(\theta_i)$,
$s_i^{(k)}(\theta_i) = \arg\max_{s_i \in S_i} E_{\theta_{-i}}[v_i(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i) + v_{-i}(K(s_i, \theta_{-i}); \theta_{-i})]$:
$0 = E_{\theta_{-i}}[\frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) +$
$+\frac{\partial v_{-i}}{\partial \kappa}(K(s_i, \theta_{-i}); \theta_{-i})\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})]$
$= E_{\theta_{-i}}[\frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) - \frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i)\frac{\partial K}{\partial s_i}(s_i, \theta_{-i})]$
$\overset{(*)}{=} E_{\theta_{-i}}[(\frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i) - \frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i))\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) +$
$+\frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i)\underbrace{(\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) - \frac{\partial K}{\partial s_i}(s_i, \theta_{-i}))}_{=0}]$.

$(*)$: we added and substracted $\frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i)\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i}))$

$(\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) - \frac{\partial K}{\partial s_i}(s_i, \theta_{-i})) = 0$ since by assumption $\frac{\partial^2 K}{\partial s_i \partial s_{-i}}(s_i, t) = 0$ (linearity of SCR) and $K(\cdot, \cdot)$ is a smooth function.

Apply Taylor expansion to the first term:
$0 = E_{\theta_{-i}}[(\frac{\partial v_i}{\partial \kappa}(K(s_i, s_{-i}^{(k-1)}(\theta_{-i})); \theta_i) - \frac{\partial v_i}{\partial \kappa}(K(s_i, \theta_{-i}); s_i))\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) =$

$= E_{\theta_{-i}}[\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i}) + \frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})(\theta_i - s_i)]\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i}))$

where $\widehat{\theta_i} \in [\min(\theta_i, s_i); \max(\theta_i, s_i)]$, and $\widehat{s_{-i}} \in [\min(s_{-i}^{(k-1)}(\theta_{-i}), \theta_{-i}); \max(s_{-i}^{(k-1)}(\theta_{-i}), \theta_{-i})]$

Since $\frac{\partial K}{\partial s_i}(s_i, s_{-i}^{(k-1)}(\theta_{-i})) \neq 0$ we get:

$$s_i - \theta_i = E_{\theta_{-i}}\left[\underbrace{\frac{\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})}}_{<0}(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i})\right],$$

$s_i =: s_i^{(k)}(\theta_i)$

We see that the distortion of type changes direction as k, the order of rationality, increases.

Moreover, from the proof of Lemma 12 we can see that
$\frac{-\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); s_i)\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}}) - \frac{\partial^2 v_{-i}}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); s_{-i})\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); s_i)} = 1,$
thus $\frac{-\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); s_i)\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); s_i)} < 1$[10]

We assume that $\hat{\theta}_i$ is close enough to s_i so that by continuity

$$\frac{-\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); \hat{\theta}_i)\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); \hat{\theta}_i)} < 1$$

as well. Take expectation of both sides:

$$E_{\theta_i}\left[s_i^{(k)}(\theta_i) - \theta_i\right] = E_{\theta_i}E_{\theta_{-i}}\left[\frac{\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})}(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i})\right]$$

as types are independent and the distributions of types coinside,

$$E_{\theta_i}\left[s_i^{(k)}(\theta_i) - \theta_i\right] = E_{\theta_{-i}}\left[(s_{-i}^{(k-1)}(\theta_{-i}) - \theta_{-i})E_{\theta_i}\frac{\frac{\partial^2 v_i}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); \widehat{\theta_i})}\right]$$

$$E_{\theta_i}\left[\left|s_i^{(k)}(\theta_i) - \theta_i\right|\right] < E_{\theta_{-i}}\left[\left|s_i^{(k-1)}(\theta_i) - \theta_i\right|\right]$$

This concludes the proof of Proposition 11 "Convergence".

[10] $\frac{-\frac{\partial^2 v_{-i}}{\partial \kappa^2}(K(s_i, \widehat{s_{-i}}); s_{-i})\frac{\partial K}{\partial s_i}(s_i, \widehat{s_{-i}})}{\frac{\partial^2 v_i}{\partial \kappa \partial \theta_i}(K(s_i, \widehat{s_{-i}}); s_i)} \in]0, 1[$

References

[1] P. Battigalli and M. Siniscalchi. Rationalizable bidding in first-price auctions* 1. *Games and Economic Behavior*, 45(1):38–72, 2003.

[2] Dirk Bergemann and Stephen Morris. Robust mechanism design. *Econometrica*, 73(6):pp. 1771–1813, 2005.

[3] Kim-Sau Chung and J. C. Ely. Foundations of dominant-strategy mechanisms. *The Review of Economic Studies*, 74(2):pp. 447–476, 2007.

[4] V.P. Crawford and N. Iriberri. Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions? *Econometrica*, 75(6):1721–1770, 2007.

[5] V.P. Crawford, T. Kugler, Z. Neeman, and A. Pauzner. Behaviorally Optimal Auction Design: Examples and Observations. *Journal of the European Economic Association*, 7(2-3):377–387, 2009.

[6] Claude d'Aspremont and Louis-Andre Gerard-Varet. Incentives and incomplete information. *Journal of Public Economics*, 11(1):25 – 45, 1979.

[7] Dorothea Kuebler and Georg Weizsaecker. Limited depth of reasoning and failure of cascade formation in the laboratory. *The Review of Economic Studies*, 71(2):pp. 425–441, 2004.

[8] Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):pp. 1313–1326, 1995.

[9] Dale O. Stahl and Paul W. Wilson. Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization*, 25(3):309 – 327, 1994.

[10] Robert Wilson. Game-thepretic approaches to trading processes. In *Advances in Economic Theory*.