ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Международный институт экономики и финансов

*Alexis V. Belianin*

# TOWARDS AN EQUILIBRIUM THEORY OF COOPERATION IN FINITELY REPEATED GAMES

Препринт WP9/2007/02

Серия WP9

Исследования по экономике и финансам
Research of economics and finance

Москва
ГУ ВШЭ
2007

УДК 330.101.542
ББК 65в6
В 41

Редактор серии WP9
«Исследования по экономике и финансам»
(Research of economics and finance)
*А.В. Белянин*

В 41 **Belianin A.V.** Towards an Equilibrium Theory of Cooperation in Finitely Repeated Games. Working paper WP9/2007/02. Moscow: State University – Higher School of Economics, 2007. – 32 p.

This paper aims at a theoretical justification to the excessive cooperative behaviour well documented in the experiments on public goods games and other settings. Cooperative strategies, while incompatible with Nash equilibrium, admit rationalization as sequentially rational strategies in dynamic games of incomplete information and partially observable actions. This point, made first by Kreps et al. (JET, 1982), remains incomplete inasmuch as beliefs which justify cooperative strategies cannot belong to the standard types space. We complete this argument by constructing an extension of the standard types space, which relaxes assumption of common knowledge of rationality, prove existence of this generalized types space, and discuss its implications for the theory of interactions under incomplete information.

*Keywords:* public goods game, prisonners' dilemma, incomplete information, universal types space, analitic sets.

*JEL codes:* C72, C92, D74.

УДК 330.101.542
ББК 65в6

Date: Ferst draft: June 2006. This draft: April 2007.

The author is thankful to Marco Novarese, Simon Gächter, Gareth Miles, Anton Suvorov and several seminar audiences for useful comments. All conclusions and errors are mine.

## 1. Introduction

Abundant experimental data show that individual behaviour often departs from the benchmark model of a selfish economic individual. This fact would probably not appear puzzling for someone who has no economics training. By contrast, qualified economists know that in the classical (finitely repeated) prisoners' dilemma, the only Nash equilibrium in strictly dominant strategies unambiguously predicts that both players should Defect, and receive Pareto-inferior payoffs. This prediction, however, fails to comply to the evidence: real people in experimental lab quite often do cooperate in a supergame when the prisoners' dilemma is repeated finitely many times (Andreoni and Miller, 1993).

Such observations are typical for many finitely repeated games, when they all contradict equilibrium predictions. One of these is the classical public goods (PG) game (Marwell and Ames (1979). Each of a group of $n \geq 2$ players is endowed with $w$ currency units, and makes an independent bid of $c_i, 0 \geq c_i \geq w$ to the common pool (public account), keeping the rest $(w - c_i)$ on her own (private account). Each unit on the private account contributes one to the utility of that *individual*, while each unit deposited on public account by any player brings $k \cdot \sum_i c_i = \alpha \bar{c}$ to the *entire group*, where $\bar{c} = \sum_i c_i / n$ is average contribution of the group and $\alpha = kn, k < 1 < kn$. Given the contributions vector $\mathbf{c}$, total utility of every individual is thus given by

$$v_i(\mathbf{c}) = w - c_i + \alpha \bar{c} \tag{1}$$

Since $1 < kn$, the efficient outcome is to contribute everything to public account. However, $k < 1$ implies that the game has a prisoner dilemma structure, and any individual is better-off depositing nothing on that account in a single-period version of this game. This result extends to any finitely repeated game, where backward induction stipulates non-cooperative behaviour in every period, which is the only subgame-perfect solution. By contrast, experimental subjects typically exhibit high or at least nonnegligible level of cooperation, especially at the early stages of the finitely repeated game.

Yet another classical example of disequilibrium behaviour is given by the trust game (Berg e.a., 1995). Player 1 receives an endowment of \$10, and decides how many dollars (\$x $\leq$ \$10) to pass to player 2. This amount is tripled by the experimenter, who gives to player 2 an amount

of 3·$x, and decides how much of that amount ($y $\leq$ 3· $x) she will return back to player 1. (All actions are observable, and payoffs are commonly known.) In another version of trust game (Kreps, 1990) two players face the following game (see Figure 1).

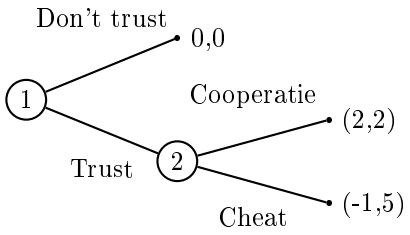


Figure 1. Trust game (Kreps, 1990).

If player 1 fails to trust the opponent, both players receive zero payoffs, while if she does, player 2 makes his choice of sharing the same pie equally (2:2) or getting the most at the expense of his trustful opponent (5:1).

Both variants of the trust game are solved by backward induction. Since player 2 strongly prefers cheating (the only equilibrium strategy) to cooperation, and player 1 knows it as well, the only (hence also subgame-perfect) equilibrium prediction is that player 1 should stop the game immediately.

However sound, this equilibrium logic of the trust game might be subject to the following criticism: If player 2 is as rational as player 1, why does not she prefer cooperation to cheating on he move? And if player 1 understands this logic, and believes player 2 will cooperate, then he should trust and give money/move to player 2. A similar logic applies to public goods game, as well as to finitely repeated prisoners' dilemma. In both cases cooperation can be sustained by pure self-interest of rational players: Inasmuch as they are *rational enough* to realize that the highest possible income would occur if everyone sticks to cooperation (until perhaps the very last stage), actions aimed at fostering this outcome do make sense, and can result in higher payoffs if these expectations are supported by others' beliefs and behaviour. This logic stands behind the decisions of many experiment participants, including those who took part

in our own recent cross-country work with Italian and Russian subjects (Belianin and Novarese, 2005)[1].

A compelling (and purely neoclassical) argument that rationalizes this intuition was first suggested by Kreps, Wilson, Milgrom and Roberts (1982) in the context of prisoners' dilemma. Their explanation makes substantial use of information incompleteness: Assume for simplicity that players can be of two types: 'rational', or *sane* with probability $\pi$, and 'irrational', or *crazy* with complementary probability $1-\pi$. Sane players are convinced by the Nash equilibrium argument, and always Hold everything on private account. Crazy players follow Tit-for-Tat strategy (Axelrod, 1981): they Contribute in period 1, and in each $t > 1$ play the same strategy that your opponent played at $t-1$. To be more specific, assume $n = 2, w = 1, c_i = \{0, 1\}, k = \frac{2}{3}$ and the discount factor is 1. The payoff matrix of a single stage game can be written (in multiples of $\frac{1}{3}$) as

| 1 \ 2 | Hold | Contribute |
|---|---|---|
| Hold | 3,3 | 5,0 |
| Contribute | 0,5 | 4,4 |

Figure 2. PG game in normal form

The only equilibrium in the stage game is (Hold, Hold), which remains subgame-perfect for any finite number $T$ of repetitions. However, if the game is repeated for $T \ll \infty$ periods, a rational player 1 who faces an opponent of unknown type can compare two strategies: Hold for all $t$, with total payoff of $5 \cdot \pi + 3 \cdot (1-\pi)(T-1)$ or Contribute for all $t$, with total payoff $0 \cdot \pi + 4 \cdot (1-\pi)(T-1)$Hence Contribute is better than Hold if $\pi < \bar{\pi} \equiv \frac{T-1}{4+T}$, which provides a rationalistic justification for cooperation for $T$ low enough. More generally, Kreps e.a. have shown that, as long as one of the players is uncertain about the type of the opponent, cooperation until the last 'few' stages (where 'few' depends on the payoffs and the probability of the opponent's rationality) is a Pareto-undominated sequential equilibrium in a finitely repeated game of incomplete information.

[1]In some of our games, degree of cooperation was even rising with time, and in one treatment even showing over 90% of cooperation on average.

While significant theoretically, this result remains incomplete in several respects. Fist, sequential equilibria are difficult to characterize explicitly (in fact, Kreps e.a. do not fully characterise one, but only determine the boundaries of a sequential equilibrium). Second, there are far too many sequential equilibria, (including noncooperative ones) and it is hard to say which of these would prevail. Third, and most importantly, this logic is incompatible with the standard properties of types space, as introduced by Harsanyi (1967-1968), and characterized by Mertens and Zamir (1985), Tan and Verlang (1988), Armbruster and Böge (1978), and others. Specifically, inasmuch as in the PG game the players do not choose strategies that survive iterated deletion of strictly dominant strategies, their rationality (in a standard sense) cannot be common knowledge[2]. and this latter is necessary for the existence of a universal types space (Brandenburger and Dekel, 1993). Hence the incomplete information explanation of cooperation along the lines of Kreps e.a. appears to be inconsistent with the existence of a canonical space of types, casting doubts of the legacy of cooperative sequential equilibria[3].

The present paper aims at filling this gap in two ways. First, starting from an explicit consideration of individual reasoning supporting cooperative behaviour in noncooperative games, we build an extended types space which fully characterizes types of each player *without* imposing common knowledge of rationality. The resulting space is based on analytic sets, rather than the usual Borel sets — hence it is more general, and also better motivated behaviourally. Despite analytic types spaces are 'richer' than the standard ones, we show that they possess the same fundamental properties, and dispense of some behaviourally controversial features of these latter. Second, we show how this extension of types space can explain a broad range of economic phenomena in which rational agents infer optimal strategies from beliefs about each others' beliefs.

---

[2]"Common certainty of rationality implies that players will choose actions that are iteratively undominated in the interim sense in the game of incomplete information." (Dekel and Gul, 1997, p.121.)

[3]Recent literature, including types spaces build in pure measure-theoretic (Heifetz and Samet, 1998) or decision-theoretic (Epstein and Wang, 1996) contexts; models of dynamic equilibria without common knowledge (Parikh and Krasucki, 1990; Heifetz, 1996), or the concept of epistemic games (Aumann and Brandenburger, 1995), all offer far-reaching opportunities of explicit incorporation of individual beliefs into equilibrium analysis, but all remain prone to the same criticism.

The paper is organized as follows. Section 2 develops the intuition behind our extension of the universal types space, and may be viewed as a non-technical summary of the argument. Section 3 develops a numerical example which illustrates the intuition behind our model in the framework of standard types space. Section 4 contains a full-fledged formal model, including construction of the analytic types space and characterizes of its basic properties. Section 5 discusses relationship of our results to the existing literature. Section 6 outlines some directions for further research and concludes. Technical proofs are collected in the Appendix.

## 2. Intuition

The logical problem with justification of cooperative beliefs in conventional types space can be illustrated using the example of Kreps e.a. (1982). Refer again to Figure 1, but now assume that *both* individuals are commonly known to be either sane or crazy, while in fact both are sane (rational). Repeating the same arguments as above, it is immediate that Contribute is better than Hold provided $\pi < \bar{\pi} \equiv \frac{T-1}{4+T}$ for both players. This logic, however, is internally flawed, because rational player 1 (symmetrically, 2) would be willing to cooperate only if she believes her opponent intends to cooperate. Since only crazy types can cooperate, agreement on cooperative strategies must imply both players know their opponents are crazy, thus it becomes common knowledge that $\pi < \bar{\pi}$ (Aumann, 1976). Yet this fact contradicts the assumption that both rationally choose to cooperate whilst been sane, i.e. $\pi \to 1$. More formally,

**Proposition 1.** *In the repeated prisoners' dilemma, cooperation cannot be achieved as equilibrium outcome if both players are sane, but believe their opponent is crazy.*

*Proof.* Let $\Omega$ be the set of states, and let the partitions (subsets of indistinguishable states) of players 1 and 2 on that set be $\mathcal{P}_1$ and $\mathcal{P}_2$. If both players agree to cooperate in the state $\omega \in \Omega$ because this will be their profit-maximizing strategy, then this state must belong to the meet (finest common coarsening) of both partitions, i.e. $\omega \in \mathcal{P}_1 \wedge \mathcal{P}_2$, and all events $E \supseteq \mathcal{P}_1 \wedge \mathcal{P}_2$ are common knowledge. In any state belonging to

such events, players believe that their opponent is crazy with high probability; given these beliefs, optimal response for a rational player is non-cooperative with probability 1 in the last stage. Reasoning backward, no player can choose cooperation, which establishes the contradiction[4]. □

The above proposition implies that, inasmuch as rational cooperation requires beliefs that the opponents are irrational, this irrationality cannot be common knowledge. In other words, to justify cooperative strategies one must allow for some sort of *rational irrationalities*, in particular, with the claim that rational players should assign positive weights only to those strategies of the opponents that are justified by any infinite sequence of such beliefs (Bernheim, 1984; Pearce, 1984). Moreover, hierarchies of mutual beliefs in the game of incomplete information form form the *types* $\Theta_i$ of all players $i = 1, \ldots N$ in the sense of Harsanyi (1967-1968). The universal types space $\Omega$ which captures both physical characteristics of the environment $S$, and players' types were effectively constructed by Mertens and Zamir (1985), Tan and Verlang (1988), Armbruster and Böge (1978), Brandenburger and Dekel (1993): the types spaces are characterized by two features:

$$\Omega = S \times \prod_{i=1}^{N} \Theta_i \tag{2}$$

$$\Theta_i = \Delta \left( S \times \prod_{j \neq i}^{N} \Theta_j \right) \tag{3}$$

The first of these characteristics claim that the universal states space is a product of physical uncertainty and all types (belief hierarchies) of each player. The second requires that the type of each player is a joint probability distribution on $S$ and the product of types of all players $j$ other than $i$. Existence of this types space requires common knowledge of players' rationality (Tan and Werlang, 1988), soits failure is incompatible with standard types space.

[4]This proof is based on the tacit assumption that rationality in the sense of expected utility maximization automatically implies reasoning by backward induction. This assumption is debatable; yet it can be shown that common knowledge of craziness remains contradictory to rationality (in the sense of expected utility maximization) even if backward induction is set aside.

## 3. The universal types space

To articulate the problems with the standard specification, consider the following simple example of discrete belief spaces. We begin with the standard types space which is defined as follows (Harsanyi, 1967-68; Mertens and Zamir, 1985). In terms of our example, suppose that both players can be sane (event S) or crazy (event C). Being exogenous, these events determine the state of physical uncertainty consisting of four elements: $X^0 \equiv \{\alpha = (1S, 1S), \beta = (1S, 2C), \gamma = (1C, 2S), \delta = (1C, 2C)\}$. In line with the above story, each rational player knows own sanity for sure, so player 1 is certain that either $\alpha$ or $\beta$ take place. He is, however, not sure of sanity of player 2, so his the space of physical uncertainty is a partition on $X^0$ given by the events that are indistinguishable for him: $X_1^0 \equiv \{\{\alpha\}, \{\beta\}\}$. In the same way, the space of physical uncertainty for sane player 2 is $X_2^0 \equiv \{\{\alpha\}, \{\gamma\}\}$.

The space of first-order uncertainty of player 1 is the set of probability distributions over his zero-level uncertainty space $X_1^0$, with generic element $p$. As prototypical space, consider $\Delta_p^1(X_1^0) = \{p_1^1(\alpha, \beta) = (1, 0), p_2^1(\alpha, \beta) = (.2, .8)\}$. Here $p_1$ and $p_2$ denote different probability distributions in $\Delta_1^1$, superscripts refer to belief levels; occasionally we shall also use subscripts to subscripts to denote particular probabilities associated with elements of $p_1^1$ etc.: thus, $p_{1_1}^1 = 1, p_{1_2}^1 = .2$ etc. Similar space for player 2 over $X^0$ is $\Delta_q^1(X_2^0) = \{q_1^1(\alpha, \gamma) = (.5, .5), q_2^1(\alpha, \gamma) = (.4, .6)\}$.

Physical uncertainties and probability distributions form spaces $X_1^0 \times \Delta_1^1(X_1^0)$ and $X_2^0 \times \Delta_2^1(X_2^0)$, which are *first-order general uncertainty spaces* with typical elements for player 1 (analogously, player 2): $(\alpha, p_1^1)$, $(\alpha, p_2^1), (\beta, p_1^1), (\beta, p_2^1)$ over player 1's beliefs, and $(\alpha, q_1^1), (\alpha, q_2^1), (\beta, q_1^1)$, $(\beta, q_1^1)$ over player 2's beliefs.

The former four elements are of no much interests to the traditional analysis, because each rational player should know his actual beliefs, so that any probability distribution which attaches positive weights to them is degenerate in every state which contains them. The latter four elements are nondegenerate — hence the second-order uncertainty of player 1 is the product of her physical uncertainty $X_1^0$, and first-order uncertainty of the other player, $\Delta_2^1(X_2^0)$, i.e. $X_1^1 \equiv X_1^0 \times \Delta_2^1(X_2^0)$. *Second-order*

*uncertainty space* of player 1 is denoted $X_1^1$, e.g.

$$\begin{aligned}\Delta_1^2(X_1^1) &= \Delta_1^2\left(X_1^0 \times \Delta_2^1 X_2^0\right)\\ &\equiv \Delta_1^2\left(\{(\alpha, q_1^1), (\alpha, q_2^1), (\beta, q_1^1), (\beta, q_2^1)\}\right)\\ &\equiv \Delta_1^2\left(\{(\alpha, (.5, .5)), (\alpha, (.4, .6)), (\beta, (.5, .5)), (\beta, (.4, .6))\}\right).\end{aligned}$$

Assume the set of probability distributions are probability distributions over that space has two elements: $p_1^2 = (.25, .25, .25, .25)$ and $p_2^2 = (.1, .1, .4, .4)$. Similar second-order uncertainty space for player 2 is

$$\begin{aligned}\Delta_2^2(X_2^1) &= \Delta_1^2\left(X_2^0 \times \Delta_1^1 X_1^0\right)\\ &\equiv \Delta_1^2\left(\{(\alpha, p_1^1), (\alpha, p_2^1), (\gamma, p_1^1), (\gamma, p_2^1)\}\right)\\ &\equiv \Delta_1^2\left(\{(\alpha, (1, 0)), (\alpha, (.2, .8)), (\gamma, (1, 0)), (\gamma, (.2, .8))\}\right).\end{aligned}$$

Let the representative elements of this space be $q_1^2 = (.2, .3, .2, .3)$ and $q_2^2 = (.1, .1, .7, .1)$. Beliefs of consecutively higher orders thus combine in sequences, which implicitly determine beliefs of the opponents over 'this' player's beliefs of lower order: thus, second-order beliefs $q_1^2$ of player 2 assign positive probabilities to all first-order beliefs of player 1. Taken to infinity, this property allows to identify all beliefs of player 1 with the set of all probability distributions over physical uncertainty and beliefs of his opponent.

Not all beliefs are equally good, however, but only those which are internally consistent. The only consistent sequence of player 2's beliefs in our example is $(q_1^1, q_1^2) \equiv ((.5, .5), (.2, .3, .2, .3))$: probabilities assigned by second-order beliefs to the zero-level states $\alpha$ and $\gamma$ are 0.5, which equal first-order beliefs over these states. By contrast, sequence $(q_2^1, q_1^2) \equiv ((.4, .6), (.2, .3, .2, .3))$ is not consistent, because $q_1^2$ assigns probability of 0.5 to the event $\alpha$, contrary to what $q_2^1$ does. Neither consistent are $(q_1^1, q_2^2) \equiv ((.5, .5), (.1, .1, .7, .1))$ and $(q_2^1, q_2^2) \equiv ((.4, .6), (.1, .1, .7, .1))$: probabilities assigned to A and B by the second-order beliefs are 0.8 and 0.2, which do not coincide with those assigned by both $q_1^1$ and $q_2^1$[5].

Going further, consider third-order uncertainty of player 1 as given by the product of her first-order uncertainty, $X_1^1 \equiv X_1^0 \times \Delta_2^1(X_2^0)$, and

[5]Most probability distributions of player 2 in our simple example are inconsistent, but this is neither upsetting nor surprising: If the set of probability distributions of any order is large enough (e.g. consists of all Borel probability measures endowed with the topology of weak convergence), *some* distributions will not be consistent, but there will be many others that will.

second-order beliefs of the other player over his second-order uncertainty, $\Delta_2^2(X_2^1)$. Let the set of distributions $\Delta_1^3$ of player 1 over her third-level space of uncertainty of order 2, $X_1^2$ be given by the following table:

| State $\setminus$ Probability | $p_1^3$ | $p_2^3$ | $p_3^3$ |
|---|---|---|---|
| $(\alpha,(.5,.5),(.2,.3,.2,.3))$ | .125 | .05 | .2 |
| $(\alpha,(.5,.5),(.1,.1,.7,.1))$ | .125 | .05 | 0 |
| $(\alpha,(.4,.6),(.2,.3,.2,.3))$ | .125 | .05 | 0 |
| $(\alpha,(.4,.6),(.1,.1,.7,.1))$ | .125 | .05 | 0 |
| $(\beta,(.5,.5),(.2,.3,.2,.3))$ | .125 | .2 | .8 |
| $(\beta,(.5,.5),(.1,.1,.7,.1))$ | .125 | .2 | 0 |
| $(\beta,(.4,.6),(.2,.3,.2,.3))$ | .125 | .2 | 0 |
| $(\beta,(.4,.6),(.1,.1,.7,.1))$ | .125 | .2 | 0 |

Distribution $p_1^3$ assigns probability $\frac{1}{4}$ to each of the four states of space of level 1, hence it is consistent with distribution $p_1^2 = (.25,.25,.25,.25)$ on $X_1^1$; however, $p_1^2$ is not consistent with $p_1^1 = (1,0)$ and $p_2^1 = (.2,.8)$ on $X_1^0$. Distribution $p_2^3$ is consistent with belies $p_2^2 = (.1,.1,.4,.4)$ on$X_1^1$ which, in turn, assigns probability 0.2 to state $\alpha$ and 0.8 to state $\beta$ on $X_1^1$, i.e. is consistent with $p_1^1$. Since $p_2^3$ also assigns $\frac{1}{5}$ to $\alpha$ and $\frac{4}{5}$ to $\beta$, the sequence of beliefs $(p_1^2, p_2^2, p_3^2)$ is consistent (at least so far). A problem with this sequence is that it assigns positive probability (of 0.75) to beliefs of player 2 other than those from the consistent sequence $(q_1^1, q_1^2) \equiv ((.5,.5),(.2,.3,.2,.3))$, i.e. those that are not consistent themselves (but which in our example do assign high probability to the event $\gamma$ of player 1 being crazy). This is not a problem for the sequence $(p_1^1, p_2^2, p_3^3)$, whose last element assigns probability 0.2 to $\alpha$ and probability 0.8 to $\beta$, as both $p_1^1$ and $p_2^2$ do, and *also* assigns probability 1 to the consistent type $(q_1^1, q_1^2) \equiv ((.5,.5),(.2,.3,.2,.3))$ of player 2.

Taken to the limit, this double consistency of the sequences of type $(p_1^1, p_2^2, p_3^3)$ are laid down in the construction of the "universal types space" as defined by Mertens-Zamir-Brandenburger-Dekel. To construct this space, extend all sequences of that last kind to infinity, and require that each player is certain in consistency of the other at any level of beliefs (this property is called *common knowledge of coherence*, see next section for details). However, it is now easy to see that this construction precludes the existence of types who would in equilibrium hold mutual beliefs the other player is crazy. Indeed, beliefs $(p_1^1, p_2^2, p_3^3)$ of player 1 assign high probability to the type of player 2 who puts high weight (of 0.6) to the event $\gamma$, while the type of player 1 consistent with this

sequence himself puts high weight (of 0.8) on the event $\beta$. Since the complementary event for both players is $\alpha$, it follows that such beliefs cannot be common knowledge. To put it another way, in order to make the Kreps e.a. argument consistent with the universal types space, we must either depart from consistence of some of the players' beliefs, or allow belief sequences of the form $(p_1^2, p_2^2, p_3^2)$, which impose high weights (of 0.8) on the event $\beta$ of player 1, while allowing for weights of 0.5 to the beliefs of player 2 in the event $\gamma$.

A bottomline of this story is now clear: in order to build the hierarchies of beliefs compatible with both the notion of types as defined by Harsanyi-Mertens-Zamir, and our observation 1, it must be either that players' beliefs are inconsistent, or that their rationality is not common knowledge[6]. Which of the two assumptions (consistency of beliefs or common knowledge of rationality) one has to abandon then? We argue that the former ought to be kept at the expense of the latter. The reason for this preference is straightforward: observation 1 agrees with the first assumption, which is basically a self-consciousness requirement, but contradicts the second one, which not only requires coordination of rationality *among* the players, but also imposes some queer restrictions on the players' rationality. Moreover, common knowledge of rationality requires that the players are able to agree on one property of each other's types — namely, consistency of each others' beliefs. This property is by far not the most transparent one; at the same time, the players fail to understand a simpler characteristics of each other, such as their beliefs over the state space. Finally, it is responsible for a number of controversial constructs, such as common prior assumption in interim types spaces (Dekel and Gul, 1997).

To proceed along these lines, we note that beliefs about each other's beliefs similar to those of $(p_1^2, p_2^2, p_3^2)$ seem to have been justifiable by the intuition of many *rational* players in cooperative experiments, such as the public goods game. A smart strategy in these games should lead to the highest possible individual payoff; yet determination of such a strategy depends of what she expects the other players to do. Both recorded and casual observation of experimental subjects (Belianin and Novarese,

[6]Other recent developments, such as introduction of particular subsets of types spaces (Ely and Pęcki, 2005; Dekel e.a., 2006) cannot do the same job.

2005) clearly suggests they explicitly build their strategies aiming at enforcement of Pareto-optimal outcome through the following observation, which gives the clue, and suggests a method to attack the problem.

**Observation 1.** *The inference of player $i$ about future actions of the opponents is based on $i$'s* perception *of beliefs of the other players, which need not the be the same as the* actual *beliefs of these other players.*

In other words, when making her inference, player $i$ would naturally think that the other players' motives and intentions are those which she (player $i$) thinks these are, which in fact they need not be. Colloquially speaking, player $i$ puts herself in place of the opponents, imposes her rationality on theirs. Notice now that observation 1 also contradicts common knowledge of rationality, as can be shown in two (rather informal) ways. First, observation 1 explicitly requires that each player $i$ must allow the other player $j$ to believe that $i$ would behave in a non-Nash way (cooperatively), as this is the condition under which $j$ himself would be willing to cooperate to maximize own payoff. Yet if $i$ believes $j$ will be cooperative, $i$'s dominant strategy is to defect in any finitely repeated version of the game, regardless of what the opponent does. Cooperation, associated with beliefs that the opponent is cooperative, is incompatible with belief in his rationality, which stipulates defection. A second way to argue is more basic, but probably also more instructive: An individual has strong incentives to cooperate if she 1) realizes the power of Nash argument, but also 2) understands that fully cooperative outcome is Pareto-superior, and 3) believes that her partners may fail to realize that defection is the only dominant-strategy equilibrium. The opponents ought then to form a mental image of our individual that would be inconsistent in a way represented by the distribution $p_2^3$ above: her beliefs of order 3 are consistent with own lower-level beliefs, yet assign strictly positive weights to inconsistent beliefs of the opponents. This constellation of rationalities is much less awkward than one may think: in fact, all individuals may be quite reasonable on their own (have consistent hierarchies of beliefs), but may mutually believe their opponents are not consistent; and it is exactly this lucky matching of incomplete rationalities which makes sustainable the cooperative outcome. This move, however, requires another logical step: a *rational* player who assumes the opponent is irrational, must admit that her beliefs over his beliefs can be arbitrary, including those which assign positive weights to his inconsistency. Inasmuch as we are still in need to build a consistent hierarchies

of beliefs about players' beliefs, hierarchies of the form $(p_1^2, p_2^2, p_3^2)$ are not sufficiently rich. To enrich this space to due depth, instead of skipping beliefs about own beliefs at the first-order interim uncertainty spaces $X_1^0 \times \Delta_1^1(X_1^0)$, we have to include these beliefs in the definition of first-order uncertainty space. With this extension, instead of $X_1^0 \times \Delta_2^1(X_2^0)$ the first-order uncertainty space becomes $X_1^0 \times \Delta_1^1(X_1^0) \times \Delta_2^1(X_2^0)$. The second-order belief space is then defined over this latter; repeat the same operation in all subsequent levels of uncertainty. This will give rise to larger hierarchies of beliefs and uncertainties: in our simplest case of two distributions at each hierarchy level, the second-order belief space shall consist not of four, but of eight elements: one for each of

$$[\alpha, (1,0), (.5,.5)], [\alpha, (1,0), (.4,.6)], [\alpha, (.2,.8), (.5,.5)], [\alpha, (.2,.8), (.4,.6)],$$
$$[\beta, (1,0), (.5,.5)], [\beta, (1,0), (.4,.6)], [\beta, (.2,.8), (.5,.5)], [\beta, (.2,.8), (.4,.6)],$$

— and so on. This extension is not complete either: such broad beliefs have to be defined for all possible physical states, including the possibility that both types are crazy, i.e. the event $\delta$. In full, this space of extended beliefs in our example takes the following form:

$$\begin{array}{cccccccc} \alpha & & \beta & & \gamma & & \delta & \\ p^1_{1_1} & & p^1_{1_2} & & p^1_{1_3} & & p^1_{1_4} & \\ p^2_{1_1} & p^2_{1_2} & p^2_{1_3} & p^2_{1_4} & p^2_{1_5} & p^2_{1_6} & p^2_{1_7} & p^2_{1_8} \\ \ldots & & & & & & & \end{array}$$

The 'real' case of Borel spaces will then lead not to the countable families $\{\Delta_i^j\}_{j=1}^{\infty}$ of countably infinite sets of all probability measures for all players $i$ but, for each player, to a single space defined as follows. To each probability distribution in a countably infinite set of all possible beliefs of order 1 there will correspond a countably infinite set of all possible beliefs of order 2; to each of these latter, a countably infinite set of all possible beliefs of order 3 etc. to infinity. By construction all subsequent beliefs form a *countably branching system* which a) contains all possible beliefs at once, b) including beliefs of consistent player 1 over inconsistent types of player 2, c) admits imposition of consistency of each player's beliefs, but d) does not impose any restriction like common knowledge of rationality. This conceptual simplification comes at a cost

of more complicated hierarchies of infinite sets which are not anymore Borel, but *analytic spaces*; hence the universal types spaces built on them shall naturally be called *analytic*. Such space is explicitly built in the next section, where we also show it does possess all the main properties of the universal types space of Mertens and Zamir (1985), while being a proper generalization of this latter.

## 4. Model

We limit attention to a class of finite games that can be viewed as games of incomplete information, in which a finite number $N$ of players make (simultaneous or sequential) decision at $T < \infty$ periods, have information (full or summary statistics) of past plays, and rationally update their beliefs about the set of possible histories (including other players' types) following these observations. A large class of games, such as trust game, ultimatum game (with unobservable types) or public goods game (a multi-stage game with imperfectly observed action) all fit into this framework, so our formulation of the model uses dynamic belief spaces formulated in a manner similar to Battigali and Siniscalchi (1997).

Define an extensive form game $\Gamma = \langle N, H, Z, \mathcal{P}, \mathfrak{I}, (u_i)_{i=1}^N, \rangle$, where $N$ is the set of players (and also, at the abuse of notation, the cardinality of that set), $H$ and $Z$ are the sets of all histories and terminal histories, respectively, $\mathfrak{I} \equiv \{\mathfrak{I}_i\}_{i=1}^N$ is the collection of all information sets of all players with typical element $I_i \in \mathfrak{I}_i$ for player $i$, $\mathcal{P}(I^t) : H^t = i (= N^t)$ is the order of moves which determines which player $i$ (set of players $N^t \subseteq N$) has to move at $t$, and $(u_i)_{i=1}^N$ are utilities (one for each player) defined on the profiles of all terminal histories.

The set of actions of player $i$ in the period $t$ will be denoted $A_i^t$, with generic element $a_i^t$; $a^t \equiv [a_1^t, \dots a_N^t]$ is the profile of actions of all players in period $t$, and $A^t \equiv \prod_{i \in N} A_i^t$ is the set of all these profiles of actions. We use $h^t \in H$ to denote an arbitrary history of the game up to period $t$, omitting the superscript index whenever time dimension is irrelevant.

Pure strategies in PG game are defined in a usual way. We use standard notations $s_i \in S_i$ etc. for arbitrary pure strategies. The set of pure strategies of player $i$ which are compatible with (do not preclude the possibility of) particular history $h^t$ are denoted $S_i^t(h^t)$, its specific element — $s_i^t(h^t)$; the set of pure strategies' profiles corresponding to specific history is $S^t(h^t) \subseteq S$, with generic element $s^t(h^t)$.

Since our motivation is largely experimental, it is worth specifying the above general framework to the case of multi-stage game with partially observable actions, as most repeated experimental games belong to that class. Information sets in these games are of the following form: for any $t \leq T$, let $[a_i^1, \ldots a_i^t]$ be any particular sequence of actions of player $i$ up to $t$. Observable histories for this player consist of the members of the sequence $[a_i^1, \ldots a_i^t]$ of player $i$, *and* of the summary statistics (amount on the public account) of choices of all players, $[\mathbf{a}_{-i}^1, \ldots, \mathbf{a}_{-i}^t]$, where $-i$ refers to all players other than $i$. Since own contributions and summary statistics are available for player's reference at any time of the game, sequences of actions observable by player $i$ up to period $t$ can be written as $\mathbf{a}_i^t \equiv [(a_i^1, \mathbf{a}_{-i}^1), \ldots, (a_i^t, \mathbf{a}_{-i}^t)]$. Let $h_i^t(\mathbf{a}_i^t)$ be any particular history up to $t$ whose possibility is *not* precluded by player $i$ who had observed sequence of actions $\mathbf{a}_i^t$ (with no further restrictions placed on the actions of the other players), and let $\mathbf{h}^t(\mathbf{a}_i^t)$ be the set of all such histories Let $u_i(\mathbf{h}^t(\mathbf{a}_i^t))$ be the set of possible payoffs to player $i$ that are not precluded after the sequence of actions $\mathbf{a}_i^t$ at the end of each period Information set of the player $i$ at the beginning of the period $t$ is then defined as $I_i^t \equiv \mathbf{h}^t(\mathbf{a}_i^t) \cup u_i(\mathbf{h}^t(\mathbf{a}_i^t))$ — i.e. the collection of past histories and expected payoffs that are not precluded by her observations[7]. Denote $\mathfrak{I}^t \equiv (I_i^t)_{i \in N}$; each strategy of player $i$ is then a sequence of maps $s_i^t(h^t) : I_i^t \to A_i^t$ from the set of histories to the set of actions feasible after each stage $t$.

Let $\mathcal{H}^t(I_i^t) \equiv \bigcup_{\tau=t}^{T} h^\tau(I_i^t)$ be the set of all continuation histories whose possibility is not precluded by the information set $(I_i^t)$. Since in experimental practice, all actions are in integer number of cents, both $2^{\mathfrak{I}^t}$ and $2^{\mathcal{H}^t}$ will be finite, and may be endowed with the natural Boolean algebras. We shall formulate the model in more general way, allowing from the outset for real-valued contributions. Accordingly, all information sets and sets of histories will be assumed to be Polish (complete separable metric), endowed with with Borel $\sigma$-algebras. The difference between the two formulations evaporates with the introduction of the set of all probability measures of the form $\mu_i^t = \mu_i^t(a_{-i}^{t+1}, a_{-i}^{t+2}, \ldots a_{-i}^T | \mathbf{h}_i^t)$ which determine players' beliefs about future actions of the other players, conditional on his observable history. All relevant spaces of these

[7]This set is generally different for different players, as different own actions are observed by different players.

measures will be denoted through $\Delta(\cdot)$ endowed with the topology of weak convergence.

Since our game is that of incomplete information, we have to consider, alongside with the space of all possible physical states $S$ (a Polish space), the set of all possible types $\Theta_i$ of all players over that set and the strategies of other players (Mertens and Zamir, 1985). Taking products, we obtain the state space $\Omega \equiv S \times \prod_{i \in N} \Theta_i$ (Brandenburger and Dekel, 1993), endowed with the product topology and the Borel $\sigma$-algebra $\mathscr{F}$ containing this latter. This space can be constructed explicitly, which construction we shall discuss in a while. At first we need to account for a dynamic nature of information flows by defining on each information set $I_i^t$ and for each player $i$ the set:

$$\mathcal{B}(I_i^t) = \left\{ B : \exists h^t \in \mathbf{h}^t(\mathbf{a}_i^t), B = S_i^t(h^t) \times \prod_{i \in N} \Theta_i \right\} \tag{4}$$

In words, $\mathcal{B}(I_i^t)$ collects all members of $\Omega$ which include those strategies whose possibility has not been precluded by the set of histories $S_i^t(h^t)$ which brought up the information set $I_i^t$. To ensure consistent sets of probability measures over $\mathcal{B}(I_i^t)$, define for each $t$ the conditional probability system (Myerson, 1986), i.e. maps of the form

$$\mu : \mathscr{F} \times \mathcal{B}(I_i^t) \to [0, 1]$$

such that

(1) $\forall B \in \mathcal{B}(I_i^t) : \mu(B|B) = 1$
(2) $\forall B \in \mathcal{B}(I_i^t) : \mu(\cdot|B)$ is a probability measure on $(\mathcal{B}(I_i^t), \mathscr{F})$
(3) $\forall A \in \mathscr{F}, B, C \in \mathcal{B}(I_i^t) : A \subset B \subset C \Rightarrow \mu(A|B)\mu(B|C) = \mu(A|C)$

The first two properties ensure that $\mu$ is a probability measure with unit mass on $\mathcal{B}(I_i^t)$. The last property says that conditional probabilities given two different subsets $B, C \in \mathcal{B}(I_i^t)$ do not contradict one another. Note that each of these conditional probabilities are conditioned upon the sequences of observable strategies $\mathbf{h}^t$ and are defined on set of continuation histories $\mathcal{H}^t$ for player $i$ through $I_i^t$.

Now let us return to the set of spaces, which naturally have to depend on histories, and incorporate individual beliefs. To build them, observe first that each conditional probability system is a subset of the set of all mappings from the set of information sets $\mathfrak{I}_i$ to the set of probability

distributions over $S \times \prod_{i \in N} \Theta_i \equiv S \times \Theta$, denoted $\Delta(S \times \Theta)$ and endowed with the topology of weak convergence. Let the subset of conditional probabilities system be denoted $\Delta^{I_i^t}(S \times \Theta)$, so that the set $X \equiv S \times \Theta \times \Delta^{I_i^t}(S \times \Theta)$ is also a Polish space endowed with product topology.

An infinite hierarchy of beliefs (Brandenburger and Dekel, 1993) can be built on these notions. At any stage of the game $t$, define recursively[8] for all players,

$$\begin{aligned} X_0^t &= S, \\ X_n^t &= X_{n-1}^t \times \prod_{i=1}^{N} \Delta X_{n-1}^t. \end{aligned} \tag{5}$$

Infinite hierarchies of beliefs of levels $n = 1, 2, \ldots$ are built on the products of spaces of the previous level and the set of all probability distributions over that space (Brandenburger and Dekel, 1993). The universal state space $\Omega^t \equiv X_0^t \times \prod_{i=1}^{N} X_\infty^t \equiv S \times \Theta$, where $\Theta \equiv \prod_i \Theta_i$, and each type $\theta_i \in \Theta_i$ of each player[9]. Individual beliefs about each other types are associated with the sequences of probability distributions of different orders $\mu_{i1}^t, \mu_{i2}^t, \ldots \mu_{in}^t \ldots$ for any $t$. Since finite products of Polish spaces, as well as the set of all probability distribution over Polish space, are all Polish, the entire space $\Omega^t$ is Polish (complete separable metric) and $\mathscr{F}^t$ the Borel $\sigma$-algebra on $\Omega^t$, which are exactly the sets and families defined above[10].

The next traditional steps are those from Brandenburger and Dekel (1993), who restrict the set of state-spaces in the following two ways:

**Assumption 1** (Consistence)**.** *Type $\theta_i \in \Theta_i$ is called* consistent *if the marginal on $x_{i1}^t$ at any levels of uncertainty over $X_0^t$ must coincide with $x_{i1}^t$, and the marginals of $x_{ij+2}^t$ on $x_{ij}^t$ must coincide with those of $x_{ij+1}^t$.*

---

[8] In what follows we focus on the general properties of the universal types space. Evolution of that space and players' beliefs appears less problematic, and is left for the future.

[9] By writing $\Theta \equiv \prod_i \Theta_i$, it is implicitly assumed that the projection of $\Theta$ on the $i^{th}$ coordinate results in degenerate distribution, i.e. each individual knows her own type.

[10] Set $\Omega^t$, being Polish (hence complete), is also compact with respect to product topology by force of the Tiknonov theorem.

This assumption says that individual beliefs of different levels do not contradict each other; since beliefs of each level are defined over beliefs of the opponents, consistence covers also beliefs of each player over those of her opponents of level $X^t_{n-1}$. This suffices to ensure homeomorphism from the set of all consistent types $\hat{\Theta}$ to the infinite hierarchy of beliefs $S \times \Theta$; this homeomorphism is denoted $f : \hat{\Theta} \to S \times \Theta$. It is important to notice, however, that these beliefs are unconstrained: each player is free to have any belief about beliefs of opponents, including those that the opponents are not consistent. The second assumption rules out these beliefs:

**Assumption 2** (Common knowledge)**.** $\forall k \geq 2, \Theta_{ik} = \{\theta_i \in \hat{\Theta} : f(\theta_i)(S \times \Theta_{ik-1}) = 1\}$.

In words, consistence must be common knowledge among all players. Denote the set of types which satisfy both assumptions at any $t$ through $\bar{\Theta}^t$, and let $\bar{\Omega}^t$ be the corresponding set of states (subset of $\hat{\Omega}^t$). This assumption suffices to show that there is homeomorphism $g : \bar{\Theta}^t \to \Delta(S \times \bar{\Theta}^t)$, which is the second main conclusion of Mertens and Zamir (1985).

Given these two assumptions, rational player should commonly choose the action which maximizes his payoffs in the entire game given his beliefs against the opponents' expected strategy, provided the opponents' types belong to $\bar{\Theta}^t$, i.e. the equilibrium strategy $s^{t*}_i \equiv [a^{1*}_i, \ldots a^{T*}_i]$ should satisfy

$$(6) \qquad s^{t*}_i = \arg\max_{s^t_i \in S^t_i} \int_{\omega^t \in \bar{\Omega}^t} \mu^t_i(\omega^t) u_i(s^t_i(\omega^t), s^t_{-i}(\omega^t), \omega^t) d(\omega^t).$$

This specification has to be relaxed under our interpretation because (observation 1) beliefs held by player $i$ in state $\omega^t$ about other players' types (and hence their actions) need not be the same as the actual beliefs which guide the opponents' actions. This simple fact precludes utilization of the above construction (and equation (6) as a criterion for optimality — see Dekel and Gul, 1997 for an extensive discussion). We propose an alternative criterion for optimization, which disposes of the common knowledge assumption of the standard framework. Alongside with the space $(\Omega^t, \mathscr{F}^t)$ of the *factual*, or *true* uncertainty, take another copy of that space, label it $(\mathcal{E}^t, \mathscr{E}^t)$, and interpret as *believed* uncertainty

of player $i$. Consider the set of all continuous images[11] $\tau_i^t$ from $(\Omega^t, \mathscr{F}^t)$ to $(\mathcal{E}^t, \mathscr{E}^t)$. These images establish correspondence between possible actual states and possible perceptions of these states by individual $i$, hence they will be called *theories* of player $i$ at $t$.

**Definition 1.** *A set $E^t \subset \mathscr{E}^t$ is called* analytic *(or A-set, or Souslin set) if it is a projection of a subset $B$ of a Polish space $\Omega^t \times \mathscr{E}^t$ onto $\mathscr{E}^t$.*

According to this definition, analytic sets emerge as members of the family of all continuous images of a Borel set whose images are located in the same space $(\mathcal{E}^t, \mathscr{E}^t)$. Any theory $\tau_{ik}^t$ of player $i$ defines an analytic set on the space of individual beliefs $(\mathcal{E}^t, \mathscr{E}^t)$, which can be identified with the subset $\Omega_A^t \subseteq \Omega^t$, again consisting of infinite hierarchies of beliefs $\Theta_A = \prod_i \Theta_{iA}$. Applying consistency requirement (1) to the state space $\Omega_A^t$, define the mapping $f : \hat{\Theta}_A \to S \times \Theta_A$ to obtain $\hat{\Omega}^t$ — the desired analytic counterpart of the universal types space.

Intuitively, the Souslin scheme extends the universal state space $\Omega^t$ in a way which (another standard property of analytic sets) is so large that an iterative application of the scheme adds nothing 'new' to that collection of sets. By contrast, a single application of the Souslin scheme to $B_{k_1,\ldots,k_n} \subset \Omega_A^t$ given by all finite collections of indices $k_1, \ldots, k_n$ already covers all possible types, and dispenses of the assumption $2$[12]. Analytic sets are relatively well studied object in descriptive set theory and general topology (see Kuratowski, 1966). In particular, these can also be defined as the sets which result from the A-operation, or the *Souslin scheme*:

$$\mathcal{A} = \bigcup_{k_1,\ldots,k_n \ldots} \bigcap_{n=1}^{\infty} B_{k_1,\ldots,k_n} \tag{7}$$

The collection of all analytic sets over $(\mathcal{E}^t, \mathscr{E}^t)$ at $t$, together with an empty set, is called *paved analytic space*, and any set $\mathcal{A} : B_{k_1,\ldots,k_n} \to$

[11] Assumption of continuity is more than mathematical convenience: *inter alia*, it implies that an event from $\mathscr{E}^t$ which player $i$ believes to be possible is indeed possible at some event in $\mathscr{F}^t$.

[12] This does not preclude, of course, the possibility of such refinements for different $\Omega^t$'s: we still may interpret $\mathcal{A}^t$ as the (evolving) system of beliefs of every players $i$ about the strategy-types pairs of all players other than $i$, which system includes the possible strategies $s_{-i}(h^t)$ of this player's opponents (not necessarily consistent with common knowledge of rationality or consistence.)

$\mathscr{E}^t$, is called *kernel* (or nucleus) of the Souslin scheme. This definition allows us to characterise analytic state space as a result of the A-operation applied for every $t$ to subsequent orders of the spaces $Y^0 =\equiv S, Y^1 \equiv Y^0 \times \Delta^1 \prod_{i=1}^N Y^0, \ldots Y^n \equiv Y^{n-1} \times \Delta^n \prod_{i=1}^N Y^{n-1}$, where each $B^t_{k_1,\ldots,k_n} \equiv \Delta^n \prod_{i=1}^N Y_i^{n-1}$. Taking $n \to \infty$, the analytic state space is just $\hat{\Omega}_A \equiv \times_{n=0}^{\infty} \Delta^n \prod_{i=1}^N Y_i^{n-1}$.

Containing by construction the standard types space of Mertens and Zamir (1985) as proper subset, the analytic types space is large enough to contain all possible beliefs, beliefs about beliefs etc., including those that provide for inconsistencies in opponents' beliefs with common knowledge of rationality, thus offering a proper types space for the argument by Kreps e.a. In particular, as analytic types do allow a reformulation of the maximization problem (6) by maximizing payoffs over analytic types spaces:

$$(8) \qquad s_i^{t*} = \arg\max_{s_i^t \in S_i^t} \int_{\omega^t \in \hat{\Omega}^t_A} \mu_i^t(\omega^t) u_i(s_i^t(\omega^t), s_{-i}^t(\omega^t), \omega^t) d(\omega^t).$$

Yet before using (8), we need to make sure that this specification makes sense, or that analytic types spaces are indeed well-defined. The answer to that question turns out to be affirmative: Analytic types spaces do indeed exist, and possess properties (2) – (3) of Harsanyi-Mertens-Zamir. Moreover, analytic types spaces can be effectively constructed in two steps. At first step, repeat the procedure used to obtain consistent types space $(\hat{\Omega}^t, \mathscr{F}^t)$. At the second, apply the Souslin scheme to that space, that is, encompass straight away *any* combination of the opponents' beliefs, beliefs about beliefs etc. contained in the same space.

**Proposition 2.** *Application of operation (1) to belief hierarchies obtained as the result of the Souslin operation results in a set* $(\hat{\Omega}^t_A, \mathscr{F}^t_A)$ *satisfying* $\hat{\Omega}^t_A = S \times \prod_{i=1}^N \Theta_{iA}$*, which is property (2) of Mertens and Zamir (1985).*

*Proof.* Any sequence $\{\mu_j\}_{j=1}^{\infty}$ of probability distribution over $(\hat{\Omega}^t, \mathscr{F}^t_A)$ partitions all lower-level distributions probabilities. Imposing consistency on the sequences of distributions, and repeating the argument of Brandenburger and Dekel (1993, p.192) yields the result through the Kolmogorov existence theorem. □

Now let us turn to the second property, (3).

**Proposition 3.** *The analytic types space* $(\hat{\Omega}_A^t, \mathscr{F}_A^t)$ *satisfies* $\Theta_{iA} \overset{hmeo}{\backsimeq} \Delta\left(S \times \prod_{j\neq i}^{N} \Theta_{jA}\right)$ *(where* $\overset{hmeo}{\backsimeq}$ *denotes homeomorphism), which is property (3) of Mertens and Zamir (1985).*

Proof of proposition 3 is technical, and contained in the Appendix.

As long as analytic sets contain Borel sets (and the inclusion is proper on any uncountable space), analytic types spaces indeed generalize the standard types spaces: they contain all types described by Mertens-Zamir (1985) *and* many more, allowing for broader set of preferences without controversial technical assumptions, such as common knowledge of rationality. This descriptive potential comes at a cost of greater technicalities, which leads to a natural question of whether the analytic types space is tractable. A positive answer to this question is contained in the following proposition:

**Theorem 1.** *For any real number* $\epsilon > 0$ *it is possible to choose analytic sets* $A, B \subset \Omega_A^t, A \cap B = \varnothing$*, such that*

- *there exists an open subset* $E \subset \Omega_A^t$*, such that*
- $A \subset E$ *and* $B \cap E = 0$*, and*
- *the distance between* $A$ *and* $B$ *is less than* $\epsilon$*.*

*Proof.* The theorem statement is essentially a variant of the regularity ($T_3$) property of topological spaces, and immediately follows from the Lusin separation theorem for analytic sets (Kuratowski, 1966, p.497). □

This proposition implies that at worst the analytic state space can be approximated by the standard (Borel) ones. Combined with theorem 3.1 in Mertens and Zamir (1985) (which shows that the standard types spaces can be approximated by finite ones), we conclude that any representation of analytic types spaces can be approximated by the finite ones to a desired degree of precision, the only difference being that this approximation is one step further.

## 5. Related literature

Let us now discuss a few implications of analytic types spaces, as well as their potential use in economic analysis. Being a proper generalization of the standard ones, they allow for greater variety of preferences, including those that do not require common knowledge of rationality, while still allowing for maximizing behaviour in in the Bayesian sense

(Tan and Werlang, 1988). *Inter alia*, this types space allows for miscoordination of rationalities (particularly, differences in depths of individual judgments about each other beliefs); and yet it is exactly this miscoordination, or failure to locate the Nash equilibrium strategy which is required to sustain Pareto-efficient outcome in the PG game. In fact, one might guess that coordination of actions (on cooperative outcome) is the consequence of this miscoordination of rationalities can be analysed in the framework of epistemic games, pioneered by Aumann and Brandenburger (1995). However, our setup is rather different from theirs, in that it allows for a much broader class of beliefs which may result in a model of player who is simultaneously rational (expected utility maximizer) and consistent (has non-contradictory beliefs), but not necessarily chooses the strategy that is equilibrium or even dominant (i.e. rationalizable).

Inasmuch as our approach it does allow for conflicting rationalities, it opens new way to incorporate a bulk of results from social sciences and humanities into game-theoretic and economics literature. This covers (but is not limited to) various aspects of human action, including social constituents of individual preferences, psychological motives for reciprocity, altruism and trust, efficiency of PR campaigns, political preferences, and even philosophical aspects of ethics, aesthetic and cognition.

The concept of analytic types space settles a theoretical platform for rigorous analysis of these issues, which seems to go beyond the potent of all existing works. Perhaps the closest in motivation to ours are works by Vladimir Lefebvre (1977, 2001). The motivating stories for his approach is an intuitive argument that human actions are often governed not by what the agents find most appropriate or reasonable to do *on their own*, but what they *think* other people will do. Moreover — and unlike most of purely game-theoretic tradition — Lefebvre explicitly articulates that when making their decisions, individual agents may and do form mental models of each others' beliefs. In terms of modern economics, this idea is pertinent to the recent literature on multiple rationalities or multiple selves, together with biases and imperfections. However, all these approaches (including Lefebvre's) are essentially *algebraic*: they consider the effect of particular attribute or belief of the agent on the optimal strategy in interaction with another individual. The power of this analysis is necessarily limited to the *listing* of possible attributes or beliefs, which listing is inevitably *closed* in some predefined sense. Our approach is different: instead of asking what kind of beliefs can arise in individual

interaction, we construct the *space* that is rich enough to contain all possible beliefs, delimited only by the very basic property of consistency, and bearing on the infinitely large set of possibilities. An important implicit assumption that underlies this construction is that mental models of the other guy are akin (in fact, homeomorphic) to mathematical concepts, such as sequences of countably branching systems. In that sense, our approach may be termed *topological*, but suggests a departure from the standard use of topological concepts. Instead of being a descriptive *tool* of analysis of an objective and independent individual (which necessarily imposes very specific and sometimes adverse structure on the space of individual beliefs), in our view, these beliefs *are* mathematical objects by their nature, and hence can and should be analyzed in their own, proper terms.

The exact potent of this approach is not clear yet, but several considerations suggest that it may be very fruitful. First, it allows to reconcile with individual rationality the fact that humans tend to deal with other people in a way that ultimately depends on the mental picture of the others. Humans tend to think of other people just as bad (or as good) as they are themselves;; and these beleifs ultimately determine the behavioural shape of the society we live in. This problem of social coordination on Pareto-preferable action is in fact implicit in the Kantian "categorical imperative" which concisely captures this problem by requiring that one ought deal with other people in the same way as we want them to deal with us. In other words, the imperative requires that we think of other people as if their motives are just like ours; and behave towards them by choosing best response to the strategy they would choose given this belief. Arguments of that sort tend to justify the long-standing judgment that apparent irrationality of individual behaviour might in fact be not only deeper than some examples of rational strategies, but be responsible for the most important examples of cooperation and coordination in those environments where traditional wisdom expects people to be entirely selfish.

Another example we take is game-theoretic. Dekel and Gul (1997) in a very deep and influential survey on rationality in game theory, argued that in games of incomplete information it is not appropriate to ground the analysis on an assumption that individual beliefs are derived from a commonly known *ex ante* model (p.114). The two reasons they site are that 1) individual rationalities may be conflicting, which precludes their

common knowledge, and 2) priors are artificial constructs, rather than actual states of the world. The concept of analytic types space immediately addresses the first of these concerns by allowing for conflicting rationalities. It also suggests a way to formally deal with the second. People in reality do ground their decisions not just on facts, but also on beliefs, dreams, and idols of all respects; and a bulk of important decisions (including whether to marry someone, whether to support particular religion, or initiate a war), is taken on purely imaginary grounds. The fact that these grounds, be these in belief, feeling or uncertainty spaces, are not physically implemented does not make them 'irreal', but just underlies the difference between natural sciences, where the object is always observable to an external observer, and economics as social science, which must deal with less tangible, but not less real matters — and the analytic types space allow exactly for that.

## 6. Conclusion

Experimental evidence suggests several conclusions whose validity seems to go beyond the laboratory. In many contexts, players seem to be able to elaborate specific coordination mechanisms which allow them to depart from the unique Nash equilibrium prediction in the PG game to their mutual satisfaction. Our analysis suggests that the main reason for this might be the failure of all players to properly assess the beliefs of each other, i.e. possible imperfection of their rationality. Generalizing from the experimental environment to more general and real-life settings, imperfect rationality of the participants of a social interaction might be the acting cause of the achievement of socially optimal outcome, which result would have been impossible had their rationality been perfect. While being rational in a weaker sense, this imperfection is in fact responsible for the possibility of coordination, and ultimately bridges the gap between individual and social rationality.

A major result of our paper is the development of an analytical framework for such an analysis. The concept of analytic types space is of course not limited to the experimental games, but seems to allow for a broad range of economic phenomena, such as cooperation, trust, reciprocity and mutual action. Why is it the case that, in some societies people are trusting each other — accept documents without verification, leave unattended belongings, give customers paybacks; while in others the same kind of actions are virtually impossible? Our model offers a line

of explanation of these phenomena in terms of different mental models underlying these behavioural strategies, again in the framework of epistemic games. All these problems seem to share the same key feature: reaching a socially desirable outcome requires particular constellation of individual beliefs and rationalities *even though* the parties may fail to realize the exact motives of each other. As such the analytic types space combine traditional game-theoretic analysis with boundedly rational behaviour. Our analysis suggest that this failure not only does not contradict maximizing behaviour, but does *de facto* make the socially profitable interaction possible at all.

This conclusion immediately raises an important question: what *are* the constellation of individual beliefs are most likely to arise in practice, *which* of these possible rationalities favour particular equilibria in dynamic interactions, and *how* individual agents 'select' the rules of thumb to use in interactions of economic relevance and interest? This question is pertinent to the issues of learning, interactive epistemology and boundedly rational behaviour, and is also most interesting from the applications' viewpoint, so we intend to address it in the near future.

## References


[1] Andreoni, James. (1995). Cooperation in Public Goods Experiments: Kindness or Confusion? American Economic Review 85(4), p.891-904.

[2] Andreoni, James, and Miller, John H. (1993). Rational cooperation in the finitely repeated prisonners' dilemma: experimental evidence. *Economic Journal*, v.103, p.570-585.

[3] Armbruster W. and Böge W. (1978). Bayesian game theory. In: O.Moeschlin and D.Pallaschke (eds.). *Game theory and related topics*. Amsterdam: North-Holland.

[4] Aumann, Robert J. and Adam Brandenburger (1995). Epistemic conditions for Nash equiliblrium. Econometrica, v.63, no.5, p.1161-1180.

[5] Axelrod, Robert (1981). The emergence of cooperation among egoists. American Political Science Review, v.75, p.306-318.

[6] Battigali, Pierpaolo, and Marciano Siniscalchi (1997). An epistemic characterization of extensive form rationalizability. Mimeo: Princeton university and Stanford university.

[7] Belianin, Alexis, and Marco Novarese. (2005). Trust, communication and equilibrium behaviour in public goods game: a cross-country experimental study. Mimeo: ICEF and University of Piedmont.

[8] Berg, Joyce; Dickhaut, John and McCabe, Kevin (1995). Trust, Reciprocity and Social History. *Games and Economic Behavior*, v.10, p.122-142.

[9] Bernheim, B.Douglas (1984). Rationalizable strategic behaviour. Econometrica, 52, 1007-1028.

[10] Brandenburger, A. and Dekel E. (1987) Rationalizability and correlated equilibria. Econometrica, v.55, no.6, p.1391-1402.

[11] Brandenburger, A. and Dekel E. (1993) Hierarchies of beliefs and common knowledge. Journal of Economic Theory, v.59, p.189-198.

[12] Camerer, Colin. (1995). Individual Decision Making. In: J.H.Kagel and A.E.Roth (eds.). *The handbook of experimental economics*. Princeton University Press.

[13] Dekel, Eddie, and Faruk Gul (1997). Rationality and knowledge in game theory. In: D.Kreps and K.Wallis, Eds. *Advances in economics and econometrics: the VIIth World Congress*. Vol.1. Cambridge: CUP.

[14] Dekel, Eddie, Drew Fudenberg, Stephen Morris. (2006). Interim correlated rationalizability. Mimeo, Princeton University.

[15] Dellacherie, Claude, and Paul-Andre Meyer. (1978). *Probability and potential*. Paris: Hermann.

[16] Ely, Richard T. and Marcin Pęcki. (2005) Hierarchies of belief and interim rationalizability. *Theoretical Economics*.

[17] Engelking, Rychard. (1982). *General topology*. Warsaw: PNW (Moscow: Nauka, 1985.

[18] Epstein, Larry and Tan Wang (1996). Beliefs about beliefs without probabilities. Econometrica, v.64, p.1343-1373.

[19] Fudenberg, Drew, and Jean Tirole (1991). *Game theory*. Cambridge: MIT Press.

[20] Güth, Werner, and Menahem Yaari (1992). Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach. In: Ulrich Witt, ed., *Explaining Process and Change. Approaches to Evolutionary Economics.* The University of Michigan Press, Ann Arbor, pp. 23-34.

[21] Harsanyi, John A. (1967-1968). Games with incomplete information played by Bayesian players. Management science, v.14, Pts.I-III, pp.159-182, 320-334, 486-502.

[22] Heifets, Aviad (1996). Comment on consensus without common knowledge. Journal of Economic Theory, v.70, p.273-277.

[23] Heifets, Aviad, and Dov Samet (1998). Topology-free typology of beliefs. Journal of Economic Theory, v.82, p.324-341.

[24] Kreps, David (1990). Corporate culture and economic theory. In: J.E.Alt and K.A.Shepsle (Eds.). *Perspectives in positive political economy.* Cambridge: CUP, p.90-143.

[25] Kreps, David, Paul Milgrom, John Roberts, Robert Wilson. (1982), Rational cooperation in the finitely repeated prisoners dilemma. Journal of Economic Theory, v.27, p.245-252.

[26] Kreps, David, and Robert Wilson. (1982). Sequential equilibria. Econometrica, v.50,p.253-279.

[27] Kuratowski, Kazimir. (1966). *Topology.* Warszawa: PWN.

[28] Levebvre, Vladimir A. (1977). Structure of Awareness: Towards a Symboloic Language Human Reflection. N.Y.: Sage.

[29] Levebvre, Vladimir A. (1980). The Algebra of Conscience. Boston: Kluwer.

[30] Marwell, Gerald, and Ruth Ames (1979). Experiments on the provision of public goods I: Resources, interest, group size and the free-rider problem. American Journal of Sociology, v.84(6), p.1335-1360.

[31] Mertens, Jean-Francois, and Shmuel Zamir. (1985) Foundations of Bayesian Analysis for Games with Incomplete Information. International Journal of Game Theory, v.14, p.1-29.

[32] Myerson, Roger. (1986). Mulitstage games with communication. Econometrica, v.54,p.323-358.

[33] Osborne, M. and Rubinstein A. (1993). *A course of game theory.* MIT Press.

[34] Parikh, R., and P.Krasucki (1990) Communication, consensus and knowledge. Journal of Economic Theory, v.52, p.178-189.

[35] Pearce, David. (1984). Rationalizable strategic behaviour and the problem of perfection. Econometrica, 52, 1029-1050.

[36] Tan T.C.F. and S.da Costa Werlang (1988). The Bayesian foundation of solution concepts of games. Journal of Economic Theory, v.45, p.370-391.

## 7. Appendix. Proof of proposition 3

We introduce first a definition and establish a number of preliminary lemmata.

**Lemma 1.** *Let $\mathcal{P} \equiv \{p \in P(X) : p(A) \geq c\}$, where $P(X)$ is the set of all probability measures on Borel measurable space $(X, \hat{\Omega}, \mathscr{F})$, and $A$ is an analytic subset of $X$ (equivalently, $Y$). The set $\mathcal{P}$ is analytic for each $c \in \mathbb{R}$*

*Proof.* The proof closely follows Bertsekas and Shreve (1978, p.166). Let $\mathcal{A}$ be the nucleus of the Souslin scheme for $\Omega^t, \mathscr{F}^t$. Let $\mathscr{N}$ be the set of irrationals (the Baire null space). For any finite collection of indices $\kappa = (\varkappa_1 \ldots \varkappa_n)$, define the sets

$$N(\varkappa_1 \ldots \varkappa_n) = \{(\zeta_1, \zeta_2, \ldots) \in \mathscr{N} : \zeta_1 = \varkappa_1, \ldots \zeta_n = \varkappa_n\} \tag{9}$$

$$\begin{aligned} M(\varkappa_1 \ldots \varkappa_n) &= \{(\zeta_1, \zeta_2, \ldots) \in \mathscr{N} : \zeta_1 \leq \varkappa_1, \ldots \zeta_n \leq \varkappa_n\} \\ &= \bigcup_{\sigma_1 \leq \varkappa_1 \ldots \sigma_n \leq \varkappa_n} N(\sigma_1 \ldots \sigma_n) \end{aligned} \tag{10}$$

Define also

$$R(\varkappa_1 \ldots \varkappa_n) = \bigcup_{z \in M(\varkappa_1 \ldots \varkappa_n)} \bigcap_{\kappa < z} \mathcal{A}(\kappa) \tag{11}$$

$$K(\varkappa_1 \ldots \varkappa_n) = \bigcup_{\sigma_1 \leq \varkappa_1 \ldots \sigma_n \leq \varkappa_n} \bigcap_{m=1}^{k} \mathcal{A}(\sigma_1 \ldots \sigma_m) \tag{12}$$

It follows that $R(\varkappa_1 \ldots \varkappa_n) \subset K(\varkappa_1 \ldots \varkappa_n)$ and $\bigcap_{n=1}^{\infty} K(\varkappa_1 \ldots \varkappa_n) \subset \mathcal{A}$ — hence each $K(\kappa)$ is closed. Let us now show that for each $c \in \mathbb{R}$, the set $\mathcal{P}$ is equivalent to

$$\{p \in P(X) : p(A) \geq c\} = \bigcap_{m=1}^{\infty} \bigcup_{z \in \mathscr{N}} \bigcap_{\kappa < z} \{p \in P(X) : p[K(\kappa)] \geq c - \frac{1}{m}\}. \tag{13}$$

Consider any measure $p'(A) \in P(X)$, and limit attention to $p'(A) \geq c$. For all such measures, there exists a sequence $(\zeta_1', \zeta_2' \ldots) \in \mathscr{N}$ such

that $p(A) = p\left(R(\zeta'_1 \dots \zeta'_n)\right) + \varepsilon$, where $p = p'$, $\varepsilon = \frac{1}{m}$. Then for all $n = 1, 2, \dots$, it holds true that

$$p'\left(K(\zeta'_1 \dots \zeta'_n)\right) \geq p'\left(R(\zeta'_1 \dots \zeta'_n)\right) \geq p'(A) - \frac{1}{m} \geq c - \frac{1}{m} \tag{14}$$

Hence $p'$ belongs to the rhs of (13). Taken now the level of hierarchies $n$ to the limit, observe that there will always exist a sequence $(\zeta_1, \zeta_2 \dots) \in \mathscr{N}$ such that

$$p'\left(\bigcap_{m=1}^{\infty} K(\zeta_1 \dots \zeta'_n)\right) = \lim_{n\to\infty} p' K(\zeta_1 \dots \zeta'_n) \geq c - \frac{1}{m} \tag{15}$$

and since $\bigcap_{n=1}^{\infty} K(\varkappa_1 \dots \varkappa_n) \subset \mathcal{A}$ is closed, it follows that $p'(A) \geq c - \frac{1}{m}, \forall m$, which establishes (13). Since the set $T_m(\kappa) = \{p \in P(X) : p[K(\kappa)] \geq c - \frac{1}{m}\}$ is Borel-measurable for all $m \geq 1$ and all sequences $\kappa$, and since $\{p \in P(X) : p(A) \geq c\} = \bigcap_{m=1}^{\infty} N(T_m)$, the sets of measures over $\mathscr{A}$ are analytic too. □

**Lemma 2.** *In the PG game, the set $\mathcal{P}$ is compact.*

*Proof.* Since in the PG game $c \in \mathbb{R}$ is bounded, any probability in $\mathcal{P}$ is a continuous image of a compact space, and hence is itself compact. □

Equipped with the measurability properties of analytic sets, let us now introduce the notion of analytic $\sigma$-algebras and analytically measurable functions. The rest of proof will be done in terms of them.

**Definition 2.** Analytic $\sigma$-algebra $\mathscr{S}$ *is the smallest $\sigma$-algebra on the set $Y$ generated by the collection of all analytic subsets of $Y$.*

**Definition 3.** *Let $A \subset \mathscr{F}_X$ and $B \subset \mathscr{E}_Y$ be arbitrary subsets of Borel spaces $X$ and $Y$ with the respective $\sigma$-algebras. A function $g : A \to B$ is called* analytically measurable *if $g^{-1}(B) \in \mathscr{F}_A$ for all $B \in \mathscr{E}_Y$.*

Set $Y = \Theta_{iA}$ and $X = \Omega_{-iA} \equiv S \times \prod_{j\neq i}^{N} \Theta_{jA}$. Now we are ready to prove the remainder of proposition 3

*Proof.* Any analytic set in $\Omega_A^t$ by construction belongs to the product of two Borel sets: $\Theta_{iA}$ and $\Omega_{-iA} \equiv S \times \prod_{j\neq i}^{N} \Theta_{jA}$. Define the projection from any $A \subset \Theta_{iA} \times \Omega_{-iA}$ to $\Theta_{iA}$ through $\rho$, and let $\rho(A) = B \subset \Theta_{iA}$ — an analytic subset of $\Theta_{iA}$. Consider an analytically measurable

mapping $g : \Theta_{iA} \to \Omega_{-iA}$, which maps an analytically measurable subset $\Theta_{iA} \in \mathscr{S}$ to the Borel space $\Omega_{-iA}$. By the Jankov-von Neumann theorem (Bertsekas and Shreve, 1978, p.182ff), for all $B \in \Theta_{iA}$, the graph of this map, i.e. the set $\{g(B), B\} \in \Theta_{iA} \times \Omega_{-iA}$ is contained in $A$. The map $g$ thus defined is a composition of a continuous map $\mathscr{N} \to \Theta_{iA} \times \Omega_{-iA}$ and a projection, hence is continuous itself.

Since the graph of a continuous map $g$ is contained in a compact set $A \subset \mathscr{A}$, it constitutes a homeomorphic embedding of $\Theta_{iA}$ in $\Theta_{iA} \times \Omega_{-iA}$ (Engelking, 1985 2.3.22, p.136). To show that $g$ is one-to-one, it suffices to notice that the space $\Theta_{iA} \equiv \times_{n=1}^{\infty} \Delta_i^n Y_i^n$ was built from all the sequences which attach some probabilities to every member of $Y_i^{n-1}$ for all $n$. Hence surjection follows from the construction of types space, and the map $g : \Theta_{iA} \to \Omega_{-iA}$ is one-to-one, onto and continuous in both directions, which establishes the desired homeomorphism. $\square$

Публикуемые в серии работы были представлены на научных семинарах, организованных в МИЭФ в рамках научной программы МИЭФ – ГУ ВШЭ, координируемой Международным академическим комитетом МИЭФ. Программа реализуется с 2003 г. при участии Директора проекта МИЭФ со стороны Лондонской школы экономики и политических наук профессора Ричарда Джекмана и старшего академического советника Амоса Витцума.

Papers published in this series were presented at the ICEF research seminars within the frame of its research programme coordinated by the International Academic Committee of ICEF. The programme has been implemented since 2003 and supervized by ICEF Project Director at LSE Professor Richard Jackman and Senior Academic Advisor Dr. Amos Witztum.

*Препринт WP9/2007/02*
*Серия WP9*
*Исследования по экономике и финансам*
*Research of economics and finance*

Alexis V. Belianin

# Towards an Equilibrium Theory of Cooperation in Finitely Repeated Games

Публикуется в авторской редакции

Зав. редакцией оперативного выпуска *А.В. Заиченко*

ЛР № 020832 от 15 октября 1993 г.
Формат 60×84 1/16. Бумага офсетная. Печать трафаретная.
Тираж 150 экз. Уч.-изд. л. 2,07. Усл. печ. л. 1,86
Заказ № . Изд. № 673

ГУ ВШЭ. 125319, Москва, Кочновский проезд, 3
Типография ГУ ВШЭ. 125319, Москва, Кочновский проезд, 3