

Doubting Others' Faultlessness and Cooperation in Centipede Games

Aviad Heifetz* and Ady Pauzner†

September 1999

Abstract

In the Centipede game, common knowledge of rationality implies the non-cooperative backward induction outcome. We consider a slight deviation from this assumption: When a player contemplates the best action at a future decision node, she assigns some small probability to the event that other players may reach a different conclusion when they carry out the same analysis. We assume that, whether players do or do not actually err, there is common certainty among them that they maintain such beliefs. We show how this implies cooperation among the players until several steps before the end of the game. Our results do not depend on forward induction or reputation effects, and thus also apply to finite horizon overlapping generations models with fiat money.

J.E.L. No.: C73. **Field:** Game Theory.

*Eitan Berglass School of Economics, Tel Aviv University, heifetz@post.tau.ac.il

†Eitan Berglass School of Economics, Tel Aviv University, pauzner@post.tau.ac.il

1 Introduction

The Centipede game (Rosenthal 1981) is a striking example of a strategic interaction in which the game theoretic solution differs considerably from what one would expect human subjects to do. In this game, continual cooperation between two parties can yield substantial gains to both, while the backward induction outcome precludes cooperation altogether.¹

Yet in many real life interactions with a similar structure, cooperation seems to be ensued a long while before it breaks down. One example is the use of state-issued fiat money. Had there been a meteorite, bound to hit and destroy Earth at the end of the next millennium, dollars would be worthless one day before the destruction. Reasoning backwards, they would be worthless one day earlier, and so forth. However, it is hard to believe that dollars would have stopped serving in trade today. The breakdown of backward induction reasoning is also observed in laboratory experiments of the Centipede game (see, for example, McKelvey and Palfrey (1992)). In these experiments, a significant proportion of subjects cooperate for several rounds.

In this paper we examine a possible deviation from the “rational” model. We assume that when a player contemplates the best action at some future decision node, she attributes some small but positive chance to the possibility that other players may reach a different conclusion when they carry out the same analysis. Moreover, we assume that each player believes that the other players bear in mind the same considerations, that they also maintain such beliefs, and so on. In other words, common knowledge of rationality is relaxed to *common certainty that with some small probability ε , the other players might conclude differently when they contemplate the best action at each decision node*.

When applied to the Centipede game, the prediction of this model differs considerably from that of the “rational” model - the backward induction reasoning induces cooperation for a long while. Cooperation breaks down only several steps before the end

¹Aumann (1995 and 1998) shows that common knowledge of rationality implies the backward induction solution.

of the game.

Importantly, our result does not rely on any “forward induction” or “reputation” reasoning (Kreps, Milgrom, Roberts, and Wilson (1982), Kreps (1990, p. 536)), in which a rational type of a player mimics her irrational, cooperative type in order to convince her opponent that she will cooperate in the future. In our model, in contrast, the action that a player takes at an early stage of the game *does not convey any information about the actions she might take later on*. In other words, our model adheres strictly to backward induction logic – the players maintain a common understanding that if a player’s decision node is ever reached, she will analyze the rest of the game as if the game starts there, and would not question “why” this node was reached.

To guarantee this feature, we focus on an overlapping generations version of the Centipede game, in which different players play in different decision nodes. Each player can secure a payoff of 1 by exiting. If she takes the risk of continuing, her payoff is 0 if the next player exits and $+d$ if the next player chooses to continue. Working with this version of the game assures that any result we obtain cannot rely on forward induction. Since each player acts only once, no other player can make inferences about her type from the history of play; thus, she has no incentive to influence others’ beliefs about her. This feature fits better some economic applications of the Centipede game. For example, in the fiat money case no agent expects her decision whether or not to accept money to affect the attitude of the rest of the population towards money.

We also analyze the two-player, alternating move version of the game (as in Rosenthal 1981). To exclude the possibility of forward induction reasoning in this setting, one has to assume that the game is played in its agent form: different agents of the same player get to play at the different decision nodes of the player. All these agents eventually enjoy the same payoff – that of the player they represent, but they make their decisions separately. This version of the game is more difficult to analyze, because there are numerous possible payoffs depending on the path of play. However, numerical analysis yields the same qualitative results as the overlapping generations version.

To get an intuitive understanding of why our model gives different predictions than the usual backward induction reasoning with full rationality, consider the following de-

liberation of player 1 at the beginning of a long Centipede game: “True, if I were in the shoes of player 2 who plays right after me, and had I assumed it is commonly known that everybody reasons like me, I would not cooperate. Thus, if player 2 reasons in this way, I should exit right away. However, in order to decide what’s best for 2, I had to put myself in the shoes of many players at consecutive decision nodes, and 2 will have to follow the same procedure. It is therefore not that unlikely that at at least some of these decision nodes, player 2 would reach a conclusion opposite to mine regarding the best action - by mistake, because her way of thinking is different, because her computational abilities are limited, or for whatever other reason. Since there are so many decision nodes where this might happen, even though with a small probability at each node, the overall probability that 2 will cooperate at the next stage may not be that small. In such a case, the best for me may be to cooperate as well.”

“In fact, if player 2 maintains similar considerations regarding the way consecutive agents reason, she might also conclude that her best action is to cooperate, *even in case she makes no mistakes, but exactly like me does not rule out the possibility of mistakes*. Thus, I should in fact ascribe a rather high probability that 2 will cooperate.”

This type of reasoning will continue to be valid as long as the players are not too close the end of the game, so that there are still enough decision nodes down the game tree in which they may doubt each other’s conclusions. Consequently, cooperation lasts with a high probability until a certain number of stages before the end. The smaller the probability ε , ascribed to the possible mismatch of conclusions, the earlier cooperation ends. In the limit of full rationality, when $\varepsilon = 0$, we obtain the no-cooperation backward induction outcome.

Our result depends crucially on the fact that in the Centipede game, a player’s best action depends on the action she believes the next player will choose: she cooperates if she is sufficiently confident that the next player will cooperate. This implies that a mistake in analyzing the next player’s action will cause a player to change her own action. This property does not hold in games such as the repeated Prisoners Dilemma. Here, players have dominant strategies at each stage. The one-shot payoff of defection is higher than that of cooperation, whether or not the other player cooperates at the

next round. Unless a player believes that her choice of action will influence the next player's move, her conclusion concerning that move is irrelevant to her choice. Thus, in the finitely repeated prisoners' dilemma, the usual backward induction outcome is robust to mistakes of the type we study.²

It is important to differentiate between mistakes in reasoning, as in our model, and mistakes in acting (trembling hands), as in Selten (1975). In a long Centipede game, mistakes in reasoning accumulate, since a mistake in figuring out the best action at any of the following nodes is enough to change the player's (perceived) best action. Hence, when there are many nodes the probability of misjudging the next player's action is large. In contrast, mistakes in acting do not accumulate, as can be easily shown by induction from the end of the game: if the correct action at some stage is to exit, there is only a small, fixed chance that the player would continue, and hence the correct action one stage earlier is, again, to exit. This means that the probability of cooperation remains bounded by the probability of a tremble, independently of the length of the game.

A number of papers have looked at different types of relaxations of the common knowledge of rationality assumption that can explain large deviations from the backward induction solution in the Centipede game. When Rosenthal (1981) introduced the Centipede game, he suggested the following possible deviation from perfect rationality: Each player might get confused when contemplating the best action in her own decision nodes (but not at the decision nodes of the other player, as in our model). The propensity to err grows with the decrease in the difference between the expected payoffs of the two possible actions of a player, and this propensity is common knowledge. Rosenthal showed, in an example, how these assumptions yield cooperation until a few steps before the end of the game.

Ben-Porath (1997) showed how common *certainty* of rationality (where the event that a player is irrational has probability 0 but nonetheless is not empty) is compatible with cooperation to some extent. This relaxation of common knowledge permits an

²By a similar reasoning, cooperation would not emerge in our agent-form setting even if a confused player in the repeated Prisoners Dilemma would play a history-dependent strategy with bounded memory (e.g. tit-for-tat). This contrasts with the reputation model of Kreps et al. (1982).

assignment of beliefs after a 0-probability event occurs, and thus can accommodate forward induction reasoning. Kreps (1990, p.561) applies the reputation model of Kreps, Milgrom, Roberts, and Wilson (1982) to the centipede game. Here, one player has an irrational type who always continues. It is shown that cooperation emerges for a long while even when the probability of the irrational type is small. Again, the result relies on forward induction: the player’s action affects the opponent’s posterior belief regarding her type. This gives the rational type an incentive to mimic the (cooperative) behavior of the irrational type. Aumann (1992) constructs examples where cooperation is sustained to a certain extent with a high ex ante probability, even though the ex ante probability of irrational types is very small, and there is a high ex ante probability that each player knows that the other is rational. Our paper differs from the above models in that forward induction reasoning is excluded. In the overlapping generations version (Section 3), each player acts only once and thus does not care what future players’ believe about her. In the alternating moves version (Section 4), we analyze the game in “agent form”. Here, the types of different agents of the same player are independent, so that the behavior of one agent does not provide information about the behavior of a future agent. Another difference is that players’ beliefs in our model are not chosen ad hoc, but rather are derived from the assumption that whenever two players analyze the same decision node, they have a small, known probability of mismatch.

The remainder of this paper is organized as follows. In section 2 we present an explicit example to illustrate the essence of the argument. The formal model for the overlapping generations version of the game is developed in section 3. Section 4 discusses the two-player, alternating moves version. Proofs are relegated to the appendix.

2 An Example

Consider the following example of an overlapping generations Centipede game. There are five players: e, d, c, b, a . At her turn, a player can secure a payoff of 1 by exiting. For all players but the last, the payoff from continuing depends on the action of the next player: it is 0 if the next player exits and +5 if the next player chooses to continue. The

last player, a , receives 0 by continuing.

The usual backward-induction argument implies that all the players choose to exit if their node is ever reached. Now suppose, instead, that every player thinks that each of the other players might get confused with a small probability, say $\varepsilon = 0.1$, independently across the decision nodes she considers in order to choose her optimal action. A confused player knows very well how to add and multiply numbers, but whenever she has to compare two payoffs x and y such that $x > y$, she concludes by mistake that y is preferred to x . Naturally, such a confused player thinks that her way of comparing payoffs is the right one. Hence, she attributes a probability of $1 - \varepsilon$ to the event that any other player who considers the same problem would think like her, and a probability of ε to the event that the other player is “confused” (according to her view) and compares payoffs in the objectively correct way.

There is common certainty among the players that they entertain these suspicions, i.e. that every one of them suspects that each of the others might, independently at each considered decision node, get “confused” (according to her view) with probability ε . For simplicity, in this section we analyze what happens when players do not actually make mistakes (yet each believes that the other players might get confused, and that they believe that she might get confused, and so on).

How do the players play the game? Let us analyze the game backwards. What does player b think that player a will do? She believes that with probability $\varepsilon = 0.1$ player a will get confused, assess 0 as greater than 1, and continue. Thus, if b continues, she expects an average payoff of

$$0.9 \times 0 + 0.1 \times 5 = 0.5$$

This is less than the 1 that she gets by quitting, and therefore she quits.

What does player c think that player b will do? She thinks that b will quit, unless either -

- 1) b understands correctly that a will quit, but gets confused and decides to continue (probability 0.1×0.9), or

2) b gets confused when she puts herself in the shoes of player a , concludes that a will quit, and given that mistake she "correctly" decides to continue (probability 0.9×0.1).

Thus, if c continues, her expected payoff is

$$0.82 \times 0 + 0.18 \times 5 = 0.9$$

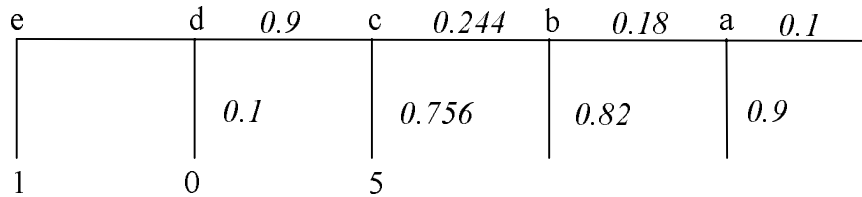
This is still less than the 1 that she can secure by quitting, and therefore she quits.

What does player d think that c will do? She believes that c quits unless she got confused exactly once when she put herself in the shoes of a , b or herself (probability $3 \times 0.1 \times 0.9^2 = 0.243$), or all the time (probability $0.1^3 = 0.001$). In the complementary event that she either never got confused or got confused two mutually-compensating times, she believes that c continues. Thus, if d continues, her expected payoff is

$$0.756 \times 0 + 0.244 \times 5 = 1.22$$

which is better than the 1 she gets by quitting. Therefore, she continues!

Finally, agent e is almost certain that d will continue. Why? If d never got confused in the shoes of a, b and c , she would continue. For the same reason, e believes that d would also continue if d got confused twice when she reasoned about a, b and c . But what if d got confused once or thrice? Also in this case d would cooperate - "by mistake", as may easily be calculated: d would then ascribe probability 0.756 that c continues. Hence, d would continue if and only if she did not get confused *in her own shoes*. Thus, e assigns a probability of 0.9 to the event that d continues. As a result, e will continue. The analysis of the game, from the perspective of e , is illustrated in the next figure (probabilities of moving in each direction are denoted by italics).



3 Overlapping Generations Centipede Games

We now turn to a general analysis of Centipede games in which there is common certainty that others players might get confused. In this section we analyze a class of Centipede games with many players, where each player gets to play only once, if and when her node is reached. As mentioned above, this class of games has several economic interpretations, and enables us to abstract from forward induction considerations such as reputation. We show how this class of games can be solved analytically. In the next section, we study the usual class of Centipede games where two players act alternately. Here, the structure is somewhat more complicated, and analytic tractability is lost. Instead, we compute numerical results for typical examples.

3.1 The Model

Consider a game with many players, where each player at her turn can either continue or quit. The player may receive one of three possible payoffs: If she quits, she receives a payoff of 1. If she continues, her payoff depends on the next player's action: if he quits, she receives 0; if he continues, her payoff is $d > 2$. The last player gets 1 by quitting and 0 by continuing.

This model fits a number of economic scenarios. Consider, for example, the use of fiat money in a world which is commonly known to end at some given future date. Agents, who live in overlapping generations, can choose between consuming their own endowment (utility of 1) or selling it in exchange for money. Each of them would enjoy a higher payoff d if the following agent accepted the money in exchange for his own endowment (hence $d - 1$ represents the gains from trade). But if the following agent declines to accept money, the utility is 0 (as paper money has no intrinsic value). If the world is known to end at a particular date, no agent would accept money at the last date. Hence, the agent playing at the one-before-last date would not give her endowment in exchange for money, since this money will be useless. Continuing this backwards induction reasoning, one can show that no agent would ever accept fiat money. However, if agents did accept money (at least most of the way to the end), almost all of them would benefit from

trade. The analysis below shows how small mutual doubts among the agents regarding each other’s maximizing behavior will induce them to use the money for trade for a long while.

The game has m decision nodes. We enumerate them *from the end* of the game. The name of each node also denotes the player who plays there. Hence, player 1 is the last one to play, 2 is the one-before-last, and so on.

Each player thinks that the next player might get confused when contemplating the optimal choice at each of the consecutive decision nodes. Thus, player $n + 1$ considers 2^n types of the other player, corresponding to whether or not he gets confused in each of the decision nodes down the game tree. We will code these types by sequences of n 0-s and 1-s, where 0 in the k -th entry corresponds to confusion at node k . As explained in the previous section, “confusion” means a reversal of the usual ordering on numbers, i.e., concluding that payoff y is preferred to payoff x when $x > y$.

In order to decide what each type would do at each of the decision nodes, we must first specify the beliefs of each type about the types of the following player. We assume that for some fixed positive $\varepsilon < \frac{1}{d}$, and independently across all decision nodes, every type assigns probability $1 - \varepsilon$ that the next player maintains the same ordering on numbers as she does at that node, and probability ε that he maintains the reverse ordering. Formally, if t_{n+1} and t_n are types of $n + 1$ and n , respectively, and the first n digits of t_{n+1} differ from the n digits of t_n at exactly ℓ entries, then t_{n+1} assigns probability $\varepsilon^\ell(1 - \varepsilon)^{n-\ell}$ to t_n . It is important to emphasize that, consequently, the types interpret their names as *neutral tags*, and do not conceive the 0-s in their names as denoting nodes where they are “mistaken”. They judge other types by taking their own type as the point of reference, and consider types very different from themselves as peculiar and rare.³

³Alternatively, we could have assumed that each type also doubts his own judgement in every node k , and assumes that both she and the consecutive player might get confused in k , independently, each with probability δ . In such a case, the player assigns probability $2\delta(1 - \delta)$ that he and the consecutive player reach opposite conclusions when contemplating the action in node k . Thus, the analysis here may be maintained with $\varepsilon = 2\delta(1 - \delta)$.

In the resulting type space there is, indeed, common certainty that each of the players suspects that the consecutive player may be confused (relative to the player's own point of view) with probability ε , independently in every node. This follows simply from the fact that all the types maintain these suspicions, and every type assigns probability 1 to the set of types of the consecutive player.

This type space admits two different interpretations. According to the interpretation that we emphasize in this paper, the type space is a hypothetical construct in the minds of the players, while the players themselves do not get confused. In other words, the actual belief of each player coincides with that of the type $111\dots 1$ which is never confused, and it is the players' mutual suspicions (which do not materialize in practice) that drive them to cooperation. An alternative interpretation is that players actually do make mistakes: ex ante each of the types may be realized, with the probability assigned by the non-confused type $111\dots 1$. In this case, the beliefs of that type represent the omniscient modeler's point of view.

Note that in either case – whether or not some types can be materialized in reality in the eyes of the modeler – the analysis of the game is the same: since every type assigns positive probabilities to all the other types of the following player, we must find out what each type would do at each decision node, in order to conclude what some type (and particularly the type $111\dots 1$) would do at previous decision nodes.

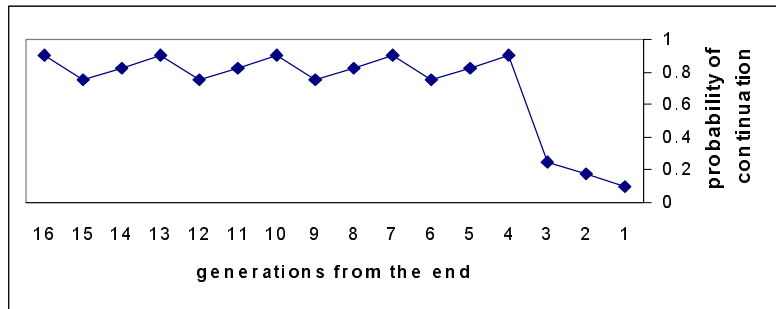
3.2 Solving the Model

How do the types decide what to do? They choose the action which maximizes their expected payoff given their beliefs, unless they are confused at their own decision node, in which case they choose the opposite action. Thus, in the case that $n+1$ is not confused at her decision node, she will choose to cooperate if she believes that the probability that n will respond by cooperation exceeds $\frac{1}{d}$: this will yield her an average payoff larger than 1, the payoff that she can guarantee by quitting immediately (the only exception is at the last node of the game tree, where the above calculation is not relevant - there player 1 simply maximizes her payoff). In the case that $n+1$ is confused, she will of course choose the opposite action to the one implied by the above rule.

Theorem 3.1. *There is an integer $n = n(d, \varepsilon)$, such that the types $111\dots 1$ of all players from n on (towards the beginning of the game) cooperate, and the types $111\dots 1$ of the last $n - 1$ players quit. $n(d, \varepsilon)$ decreases in both d and ε . When going backwards from the last node to the first, the probability that $111\dots 1$ ascribes to cooperation from the consecutive player:*

- 1) *Starts with ε at node 1, and increases monotonically until node n , which is the first where it surpasses $\frac{1}{d}$;*
- 2) *Jumps to $1 - \varepsilon$, from which it decreases monotonically until node $2n$, which is the first node where it decreases below $1 - \frac{1}{d}$;*
- 3) *The sequence of probabilities in 2) repeats itself every n nodes, as long as there are nodes in the game.*

The following graph depicts this result for the parameters of the example in the previous section ($d = 5$, $\varepsilon = 0.1$) -



The intuition is as follows. In the eyes of the type 11 of player 2, player 1 (that plays at the last node) cooperates only if she is confused – which has a probability of ε . In the eyes of the type 111 of player 3, player 2 cooperates if he got confused either at the last node and not at his own, or vice versa – altogether with probability $2\varepsilon(1 - \varepsilon)^4$. Similarly, at the few preceding nodes, the type $111\dots 1$ of a player believes that the consecutive player cooperates if he got confused an odd number of times at these nodes. As the

⁴For further elaboration see the example in the previous section.

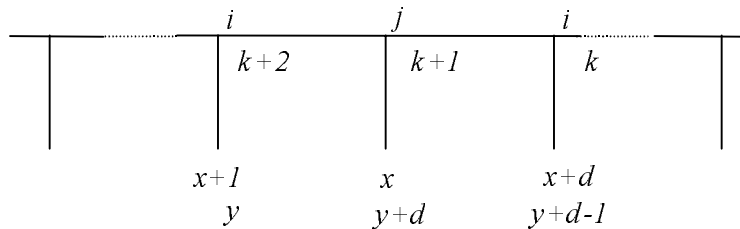
distance from the end increases, the probability of getting confused an odd number of times increases towards half. Denote by n the stage in which it surpasses $\frac{1}{d}$. At that stage, the $111\dots 1$ type would cooperate, since the probability that the next player continues is sufficiently high to compensate for the loss of 1 in case the next player exits. However, it is not only the $111\dots 1$ type who cooperates; every type who got confused an *even* number of times cooperates, since the probability he assigns to cooperation at the next stage coincides with the one assigned by the $111\dots 1$ type. This means that *all* the types of player $n + 1$ assign a high enough probability to cooperator types of player n , and hence cooperate, provided that they don't get confused at node $n + 1$ itself. At that node, therefore, the probability of cooperation (in the eyes of type $111\dots 1$ of player $n + 2$) jumps to $1 - \varepsilon$.

In node $n + 2$ from the end, the player cooperates unless he got confused either in node $n + 1$ but not in his own, or vice versa. The probability of quitting is thus $2\varepsilon(1 - \varepsilon)$. Similarly, in the few preceding nodes the agent quits if he got confused an odd number of times in the nodes starting from $n + 1$. Going backwards, in node $n + n$ this probability exceeds $\frac{1}{d}$ for the first time after $n + 1$. In the eyes of the types that quit in nodes $n + 1$ to $n + n$, who estimate the probabilities reversely (they think that types similar to themselves are common!), $n + n$ is the first time that the probability of *cooperating* exceeds $\frac{1}{d}$. Thus, at stage $n + n + 1$ each confused type says to herself: "I know that it's good to quit. However, at $n + n$ there are already enough dummies who mistakenly think that the best option is to cooperate. Therefore, I should cooperate as well."

As a result, in node $n + n + 1$ from the end these types again join the others in cooperating, unless they get confused in that very node. Hence, the probability of cooperation in $n + n + 1$ (in the eyes of type $111\dots 1$ of player $n + n + 2$) jumps again to $1 - \varepsilon$, and the cycle described in the previous paragraph repeats itself.

4 Centipede Games with Two players

Consider now the usual class of Centipede games with two players, i and j , who alternate in moves along m consecutive decision nodes. At each node, the player (say i) chooses whether to quit or to continue. If she quits, the game ends and she receives some payoff, say $x + 1$. If she continues, her payoff depends on j 's choice: if j quits, i gets x , but if j cooperates and then i quits, i gets $x + d$, where the increment $d > 2$ is a parameter of the game. If the player in the last decision node cooperates, the payoffs to the players are as if the other player had to play again and decided to quit.



Each player thinks that the other might get confused when contemplating the optimal choice at each of her decision nodes. Thus, each player considers 2^m types of the other player, corresponding to whether or not he gets confused in each of the decision nodes of the game tree.⁵ We will code these types by sequences of m 0-s and 1-s, where 0 in the k -th entry corresponds to confusion at node k from the end. As explained in the previous section, “confusion” means a reversal of the usual ordering on numbers, i.e. concluding that payoff y is preferred to payoff x when $x > y$.

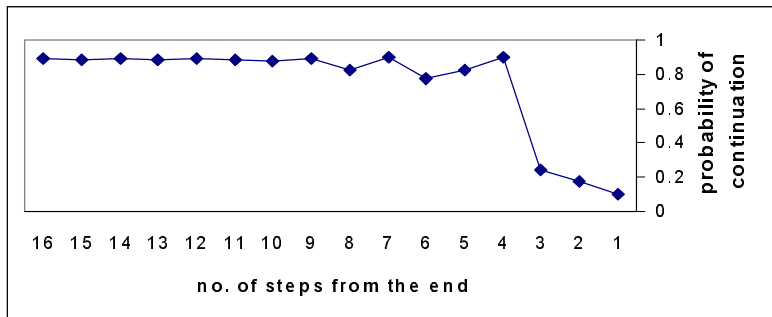
The beliefs of each type about the types of the other player are determined as in the overlapping generations case. We assume that for some positive $\varepsilon < \frac{1}{d}$, and independently across all decision nodes, every type assigns probability $1 - \varepsilon$ to the event that the other player maintains the same ordering on numbers as she does in that node, and probability ε that he maintains the reverse ordering. Formally, if t_i and t_j are types

⁵In fact, the first player may consider only 2^{m-1} types of the second player, corresponding to the $m - 1$ nodes that follow the first node, and similarly the second player may consider only 2^{m-2} types of the first player, corresponding to the $m - 2$ nodes that follow the second node, in which he chooses for the first time. This subtlety is of no consequence for the sequel.

(sequences of m 0-s and 1-s) of i and j , respectively, which differ from each other in ℓ of the m entries, then each type assigns probability $\varepsilon^\ell(1 - \varepsilon)^{m-\ell}$ to the other.

The actions of each type in each node are determined by backward induction. The player in node 1 (the last node) maximizes her payoff and quits if her type starts with 1, but if her type starts with 0 she takes the opposite action and cooperates. Inductively, the player at node k from the end takes the action that maximizes her expected payoff given her beliefs if the k -th digit if her type is 1, but if this digit is 0 she takes the opposite action.

Because in this game a player's payoff from cooperation has many possibilities, depending on the extent of cooperation, the analytic tractability of the overlapping generations case is lost. Hence, we can only obtain numerical results. The following graph depicts these results for the parameters $d = 5$, $\varepsilon = 0.1$ as in the example in section 2:



Qualitatively, the main feature of the results of the overlapping generation model are maintained: except for the last three rounds, the players cooperate with high probability. The cycles described above appear, but with decreasing length and amplitude the further the players are from the end of the game. The reason for this difference is that a type's action does not depend only on the parity of the number of mistakes he has made. Types with the same parity may disagree about the extent of future cooperation, and hence about their expected payoff. Thus, it is not necessarily true that they all switch simultaneously to cooperation exactly at the same node.

A Proofs

To prove theorem 3.1 we first need some definitions. Consider the following two complementary subsets of the set of types T_j of player j :

$$E_j^{\ell,m} = \{\text{the types of } j \text{ which are confused an even number of times in nodes } \ell, \dots, m\}$$

$$O_j^{\ell,m} = \{\text{the types of } j \text{ which are confused an odd number of times in nodes } \ell, \dots, m\}$$

The following two lemmata will be useful for the sequel.

Lemma A.1. *For $i = j + 1$, all the types in $E_i^{\ell,m}$ assign the same probability to $E_j^{\ell,m}$, all the types in $O_i^{\ell,m}$ assign the same probability to $O_j^{\ell,m}$, and these two probabilities are the same.*

Proof. For every type $t_i \in T_i$, denote by P_{t_i} the permutation of T_j , that for each $t_j \in T_j$ offsets 0-s to 1-s and 1-s to 0-s in all the entries of t_j between 1 and j in which t_i is confused (has 0 entries). When $t_i \in E_i^{\ell,m}$, the number of these entries between ℓ and m is even, so P_{t_i} maps $E_j^{\ell,m}$ to itself and $O_j^{\ell,m}$ to itself. Denote now by $\mathbf{1}_i$ the type 111...1 of player i , and identify the types with their beliefs. Then by definition $t_i = \mathbf{1}_i \circ P_{t_i}$, and in particular

$$t_i(E_j^{\ell,m}) = \mathbf{1}_i \circ P_{t_i}(E_j^{\ell,m}) = \mathbf{1}_i(E_j^{\ell,m}), \quad t_i(O_j^{\ell,m}) = \mathbf{1}_i \circ P_{t_i}(O_j^{\ell,m}) = \mathbf{1}_i(O_j^{\ell,m})$$

Similarly, for every type t'_i in $O_i^{\ell,m}$, the permutation $P_{t'_i}$ sends $E_j^{\ell,m}$ to $O_j^{\ell,m}$ and vice versa. Hence for all the types t'_i in $O_i^{\ell,m}$

$$t'_i(E_j^{\ell,m}) = \mathbf{1}_i \circ P_{t'_i}(E_j^{\ell,m}) = \mathbf{1}_i(O_j^{\ell,m}), \quad t'_i(O_j^{\ell,m}) = \mathbf{1}_i \circ P_{t'_i}(O_j^{\ell,m}) = \mathbf{1}_i(E_j^{\ell,m})$$

Thus for every t_i in $E_i^{\ell,m}$ and t'_i in $O_i^{\ell,m}$

$$t_i(E_j^{\ell,m}) = t'_i(O_j^{\ell,m})$$

as required. ■

In what follows, “the action of type t_j (of player j) in node m ” means the action of player m whose type is the m -initial segment of t_j .

Lemma A.2. *Suppose that all the types of j in $E_j^{\ell,m}$ take the same action in node m (from the end), and all the types in $O_j^{\ell,m}$ take the opposite action in that node. Denote $p = \mathbf{1}_i(E_j^{\ell,m})$.*

- 1) *If $\frac{1}{d} < p < 1 - \frac{1}{d}$, then all the types of i with 1 in entry $m + 1$ cooperate in node $m + 1$, and all the types of i with 0 in entry $m + 1$ quit in that node;*
- 2) *Otherwise, all the types of i in $E_i^{\ell,m+1}$ take the same action in node $m + 1$, and all the types of i in $O_i^{\ell,m+1}$ take the opposite action in that node.*

Proof. In case 1), all the types of i assign probability more than $\frac{1}{d}$ that j cooperates in node m . This is because if the types in $E_j^{\ell,m}$ cooperate, then by lemma A.1 all the types in $E_i^{\ell,m}$ assign to $E_j^{\ell,m}$ probability $p > \frac{1}{d}$, and the other types of i , those in $O_i^{\ell,m}$, assign to $E_j^{\ell,m}$ probability $1 - p > \frac{1}{d}$. Similarly, if the types in $O_j^{\ell,m}$ cooperate, all the types in $E_i^{\ell,m}$ assign to $E_j^{\ell,m}$ probability $1 - p > \frac{1}{d}$, and the types in $O_i^{\ell,m}$ assign to $E_j^{\ell,m}$ probability $p > \frac{1}{d}$.

Thus, all the types of i who are not confused in node $m + 1$ from the end cooperate, and all those who are confused there quit.

In case 2), denote by C_j^m the set of types of j that cooperate in node m from the end – $E_j^{\ell,m}$ or $O_j^{\ell,m}$. By lemma A.1, the types of i in one of the sets $E_i^{\ell,m}, O_i^{\ell,m}$ assign to C_j^m a probability smaller than $\frac{1}{d}$, and the types of i in the other set assign to C_j^m a probability bigger than $1 - \frac{1}{d}$. Denote the former set by Q_i^m and the latter by C_i^m .

To maximize expected utility, the types of C_i^m who are not confused in node $m + 1$ cooperate there, and the types of C_i^m who are confused in node $m + 1$ quit. In contrast, the types of Q_i^m who are not confused in node $m + 1$ quit there, and the types of C_i^m who are confused in node $m + 1$ cooperate. So altogether, the types of i who cooperate in node $m + 1$ constitute one of the sets $E_i^{\ell,m+1}, O_i^{\ell,m+1}$, and the types who quit there constitute the other set. ■

Proof of Theorem 3.1. As usual, we label the nodes from the end backwards. In node 1 (the last), only the types who are confused there cooperate. Their probability is ε . Taking $\ell = m = 1$, we are in case 2) of lemma A.2, with $p = 1 - \varepsilon > 1 - \frac{1}{d}$, which

implies that with $\ell = 1$, $m = 2$ the premise of lemma A.2 is fulfilled. We can thus iteratively apply lemma A.2 with $\ell = 1$ and increasing m , until we hit case 1). This indeed happens, because the probability p of getting confused an even number of times decreases and tends to half. Let n be the minimal m for which $p < 1 - \frac{1}{d}$. Evidently, the smaller the ε and the smaller d , the larger is n . With $\ell = 1$ and $m = n$ in the premise of lemma A.2, we are in case 1).

The lemma now implies that the premise of the lemma holds for $\ell = m = n + 1$ with $p = 1 - \varepsilon$. Keeping $\ell = n + 1$, and taking $m = n + 2$, we are back to case 2). Again we can iteratively apply lemma A.2 with $\ell = n + 1$ and increasing m , until we hit case 1). This takes exactly n nodes as above. For $\ell = n + 1$ and $m = 2n$ we are in case 1).

The procedure in the previous paragraph can be now repeated again and again, with ℓ and m increased by another n in each round, until we reach the beginning of the game tree, in the middle or at the end of one of the rounds. ■

References

- [1] Aumann, R.J. 1992. Irrationality in Game Theory, in Dasgupta et al. (eds.), *Economic Analysis of Markets and Games, Essays in Honor of Frank Hahn*, MIT Press, Cambridge.
- [2] Aumann, R.J. (1995). "Backward Induction and Common Knowledge of Rationality" *Games and Economic Behavior* 8:6-19.
- [3] Aumann, R.J. 1998. "On the Centipede Game." *Games and Economic Behavior* 23:97-105.
- [4] Ben-Porath, E. 1997. "Rationality, Nash Equilibrium and Backward Induction in Perfect-Information Games." *Review of Economic Studies* 64:23-46.
- [5] McKelvey, R.D. and T.R. Palfrey 1992. "An Experimental Study of the Centipede Game." *Econometrica* 60:803-836.
- [6] Kreps, D. 1990. *A Course in Microeconomic Theory*, Princeton University Press.
- [7] Kreps, D., P. Milgrom, J. Roberts and R. Wilson 1982. "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma." *Journal of Economic Theory* 27:245-252.
- [8] Rosenthal, R.W. 1981. "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox." *Journal of Economic Theory* 25:92-100.
- [9] Selten, R. 1982. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Form Games." *International journal of Game Theory* 4:25-55.